CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Publications from Karolinska Institutet

From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Quantifying cancer patient survival; extensions and applications of cure models and life expectancy estimation

Therese M-L Andersson

Stockholm 2013

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet. Printed by Universitetsservice AB.

© Therese M-L Andersson, 2013
ISBN 978-91-7549-297-1

Abstract

Cancer patient survival is the single most important measure of cancer patient care. By quantifying cancer patient survival in different ways further insights can be gained in terms of temporal trends and differences in cancer patient survival between groups. The objective of this thesis is to develop and apply methods for estimating the cure proportion and loss in expectation of life for cancer patients.

In paper I, a cure model was used to study temporal trends in survival of patients with acute myeloid leukaemia in Sweden. Cancer patient survival was estimated in a relative survival setting and quantified as the proportion cured and the median survival time of uncured for different age groups and by calendar time of diagnosis. We found a dramatic increase in the cure proportion for the age group 19-40, although almost no improvement was seen for patients aged 70-79 at diagnosis.

In paper II, a flexible parametric cure model was developed to overcome some limitations with standard parametric cure models. This model is a special case of a non-mixture cure model, using splines instead of a parametric distribution for the modeling. The fit of the flexible parametric cure model was compared to the fit of a Weibull non-mixture cure model, and shown to be superior in cases when the standard non-mixture cure model did not give a good fit or did not converge. Software was developed to enable use of the method.

In paper III, the possibility of using a flexible parametric relative survival model for estimating life expectancy and loss in expectation of life was evaluated. Extrapolation of the survival function is generally needed, and the flexible parametric relative survival model was shown to extrapolate the survival very well. The method was evaluated by comparing survival functions extrapolated from 10 years past diagnosis to observed survival by the use of data with 40 years of follow-up. Software was developed to enable use of the method.

In paper IV, the life expectancy and loss in expectation of life was estimated for colon cancer patients in Sweden. Even though relative survival was similar across age for colon cancer patients, the loss in expectation of life varied greatly by age, since young patients have more years to lose. We also found that the life expectancy of colon cancer patients improved over time. However, the improvement has to a large extent mimicked the improvement seen in the general population, and therefore there were no large changes in the loss in expectation of life.

In conclusion, the methods presented in this thesis are additional tools for estimating and quantifying population-based cancer patient survival, that can lead to an improved understanding of different aspects of the prognosis of cancer patients.

List of publications

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals.

I. Andersson TM, Lambert PC, Derolf AR, Kristinsson SY, Eloranta S, Landgren O, Björkholm M, Dickman PW
Temporal trends in the proportion cured among adults diagnosed with acute myeloid leukaemia in Sweden 1973–2001, a population-based study.
Br J Haematol. 2010 Mar;148(6):918-24. doi: 10.1111/j.1365-2141.2009.08026.x.

II. Andersson TM, Dickman PW, Eloranta S, Lambert PC
Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models.
BMC Med Res Methodol. 2011 Jun;11(1):96. doi: 10.1186/1471-2288-11-96.

III. Andersson TM, Dickman PW, Eloranta S, Lambe M, Lambert PC
Estimating the loss in expectation of life due to cancer using flexible parametric survival models.
Stat Med. Article first published online: 23 AUG 2013. doi: 10.1002/sim.5943

IV. Andersson TM, Dickman PW, Eloranta S, Sjövall A, Lambe M, Lambert PC
The loss in expectation of life after colon cancer: a population-based study
Manuscript

Other relevant publications:

- Andersson TM, Lambert PC
Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models.
The Stata Journal. 2012;12(4):623-628.

Contents

List of abbreviations

AIC	Akaike information criterion
AML	Acute myeloid leukaemia
BIC	Bayesian information criterion
DCO	Death certificate only
HMD	Human Mortality Database
ICD-7	International Classification of Diseases, Revision 7
K-M	Kaplan-Meier
LE	Life expectancy
LEL	Loss in expectation of life
LR	Likelihood ratio
PELL	Proportion of expected life lost
RSR	Relative survival ratio

1 Introduction

Data from cancer registries are used to measure the incidence of cancer, cancer mortality and the prognosis of cancer patients. Cancer patient survival is the single most important measure of cancer patient care (the diagnosis and treatment of cancer) and is of considerable interest to clinicians, patients and researchers. There is, however, room for improvement to the statistical methods currently used to estimate cancer patient survival. Cancer patient survival is often reported as the 5-year relative survival, the proportion of patients still alive 5 years after diagnosis if cancer was the only possible cause of death [1]. Even though there are alternative less simplistic measures, of more relevance to patients and clinicians, they have not yet been widely used.

Relative survival is defined as the ratio of the observed survival to the expected survival. Excess mortality is the mortality analogue of relative survival. It is often of interest to model the excess mortality rate to compare different groups. This is often done using Poisson regression and results are presented as excess hazard ratios [2]. This is a relative measure and does not provide an estimate of absolute risk. The popularity of presenting hazard ratios, from Poisson regression (or Cox regression for cause-specific survival) has led to researchers thinking less about absolute measures of risk and therefore the impact a potential risk factor has on both the individual and the population. The flexible parametric survival model [3, 4, 5] is an alternative model that models rates in an improved way, by directly modelling the effect of time using restricted cubic splines [6, 7]. In addition, time-dependent effects can easily be incorporated and the models also enable the results to be displayed in a number of ways that aid the understanding of individual risk and the impact on the whole population. Flexible parametric survival models have been extended to a relative survival framework, allowing the modelling of excess mortality [8].

This PhD project focuses on estimation of cancer patient survival in a population-based setting. I have extended the flexible parametric survival model to enable estimation of alternative measures of cancer patient survival, namely cure and loss in expectation of life. An important task for biostatisticians is to communicate the results from complicated statistical methods in a way that is understandable for other researchers, clinicians and healthcare policymakers. Therefore, I have also applied novel methods for estimation of cancer patient survival to cancer registry data to present the methodology to a broader audience.

This thesis combines the development of statistical methods for population-based cancer survival studies, development of user-friendly software to facilitate wide application of the methods as well as applications of methods for population-based cancer survival studies to data from cancer registries.

2 Background

2.1 What is cancer?

Cancer is not one disease, but a broad group of about 200 different diseases that all arise because of uncontrolled growth of cells in the body. Different types of cancer have different etiology, symptoms and prognosis. Cancer cells invade the tissue around the tumour and often have the ability to spread to other parts of the body. When the cancer has spread and developed tumours at other locations in the body these are referred to as metastasis, as opposed to the original tumour, which is often referred to as a primary tumour. The cells in the metastatic, or secondary, tumours are similar to those in the primary tumour, and it is the origin of the primary tumour that determines the type of cancer. For example, if a breast cancer metastasises to the lung, the tumours in the lung are metastatic breast cancer and not lung cancer.

In 2008 more than 12 million cancer cases occurred worldwide, and 7.6 million cancer deaths [9]. Cancer is the leading cause of death in many economically developed countries, and the second leading cause of death in many other countries, including many economically developing countries. The most common cancer among females worldwide is breast cancer, and it is also the cancer leading to most deaths. Among males the most common cancers are lung (the most common in developing countries) and prostate cancer (the most common in developed countries), and lung cancer is the cancer leading to most deaths. In Sweden, cancer is the second most common cause of death after cardiovascular diseases [10]. In 2011, almost 58 000 cancers were diagnosed in Sweden, and almost 23 000 deaths were attributed to cancer [11]. The most common cancers in Sweden are prostate cancer (9 663 cases in 2011) and breast cancer (8 427 cases). Other common cancers are non-melanoma skin cancer, colon cancer and lung cancer.

2.2 Cancer registries

The task of population-based cancer registries is to collect and store information on all diagnosed cases of cancer in a region or a country as well as producing statistics of the occurrence of cancer and the survival of cancer patients. They play an important role in analyzing the impact of cancer and the cancer care in the society, and how it changes over time [12, 13, 14]. There are three commonly reported outcome measures estimated from cancer registry data; incidence, mortality and survival [1]. Cancer incidence is a measure of how frequently the cancer is diagnosed and is often measured as the number of new cases per 100 000 person-years. Cancer incidence is an important measure of the cancer burden.

Cancer mortality measures deaths due to cancer in the whole population and is also given as a rate (often per 100 000 person-years). Cancer mortality provides another measure of the cancer burden, and is often presented together with the incidence to give a more complete picture of the cancer burden. Since cancer mortality reflects changes in cancer incidence as well as changes in cancer survival there are difficulties in using it to analyse cancer care. Moreover, both the incidence and the mortality are heavily influenced by the distribution of risk factors in the population, whereas the survival is not affected by the distribution of risk factors to the same extent. Improvements in cancer care, the diagnosis and treatment of cancer, are best monitored using the survival among those diagnosed with cancer.

In this thesis data from the Swedish Cancer Registry (paper I, III and IV) and the Finnish Cancer Registry (paper II) were analysed. More information about the registries is given in Section 5.

2.3 Cancer patient survival

Cancer patient survival is measured by the time between diagnosis and death due to cancer, and is often summarised as the proportion of patients surviving the cancer up to a certain point, often 5 years, after diagnosis. Cancer patient survival often aims at estimating net survival [1], sometimes called marginal survival [15]. Net survival is interpreted as the proportion of patients that survive in the hypothetical scenario where the cancer of interest is the only possible cause of death. So, the 5-year survival is an estimate of the proportion of patients still alive five years after diagnosis if the cancer of interest is the only possible cause of death. For most types of cancer 5-year survival has increased over the last few decades, indicating that cancer treatment has improved. Different research questions, and different consumers of cancer patient survival statistics, need different measures of survival, since there is no "one-size-fits-all" measure. Therefore, it is important that alternatives to the 5-year survival proportion are considered. One alternative measure is the cure proportion, the proportion of patients that will not experience excess mortality due to the diagnosed cancer. Another measure is the life expectancy among cancer patients and the loss in life expectancy due to the diagnosed cancer. Estimation of cure and the loss in expectation of life is the main focus of the studies included in this thesis, and will be described in detail in the following sections.

One problem with survival is that it can be affected by lead-time bias. Lead-time is survival time that is added to a patient's survival time because of an earlier diagnosis irrespective of a possibly postponed time of death. In the presence of lead-time, the survival time of the patients is prolonged by an earlier diagnosis, and the survival proportion at any given time point is therefore increased even if no real improvement in survival is experi-

enced. This is referred to as lead-time bias. When a cancer is diagnosed earlier, there is usually a better chance of surviving the cancer or postponing death, which will also lead to an improved survival. Therefore, it can be difficult to disentangle how much of an observed improvement is due to lead-time and how much is due to an actual improvement in survival. The possibility of lead-time should always be taken into account when comparing survival proportions between groups or over calendar time, especially for cancers that can be detected by screening. This is one of the reasons why it is very difficult to evaluate the effectiveness of screening programmes in improving cancer patient survival. Even so, cancer patient survival is the principal measure of the effectiveness of cancer care, and all studies in this thesis concern measures of cancer patient survival.

2.4 Survival analysis

All of the studies in this thesis concern the estimation of cancer patient survival, and before describing the methods used, a brief introduction to survival analysis is given in this chapter. Survival analysis is used to study the time, T, to occurrence of some event of interest [16]. In cancer patient survival that refers to time from diagnosis of cancer until death due to cancer.

It is usually not possible to follow all patients until the event of interest (death due to cancer) occurs, and when this is the case the survival time is censored at the last available follow-up time. This type of censoring, when the event is known to have not happened at last follow-up time, is referred to as right-censoring. There are different types of censoring, but in population-based cancer patient survival right-censoring is most common, and is the only type of censoring considered in this thesis. To distinguish between event times and censoring times all patients have, in addition to the observed time t_i, an event indicator d_i that takes the value 1 if the patient experienced the event and 0 if censored.

In addition to the density function, $f(t)$, and the cumulative distribution function, $F(t)$, there are three important functions to describe T. These are the survival function, the hazard function and the cumulative hazard function. The survival function gives the probability that the survival time is greater than time t, which is the same as 1 minus the cumulative distribution function

$$S(t) = P(T > t) = 1 - F(t). \tag{1}$$

The survival function is non-increasing and can only take values between 0 and 1. The hazard function is defined as the event rate at time t conditional on surviving up until time t,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2}$$

The cumulative hazard function is the integrated hazard function,

$$H(t) = \int_0^t h(u)du. \tag{3}$$

Although hard to interpret in itself, the cumulative hazard function is a useful function, for example when modelling. In the flexible parametric survival model (described in section 2.6) modelling is performed on the log cumulative hazard scale.

All the functions above are mathematically related, and some important relationships are presented here.

$$S(t) = \exp(-H(t)) = \exp\left[-\int_0^t h(u)du\right] \tag{4}$$

$$h(t) = \frac{d}{dt}H(t) = f(t)/S(t) = -d\ln(S(t))/dt = -S'(t)/S(t). \tag{5}$$

2.5 Relative survival

The method of choice for estimating cancer patient survival in a population-based setting is relative survival, $R(t)$ [1, 17], and the most common summary measure for cancer patient survival is the 5-year relative survival ratio (RSR). Relative survival is defined as the observed (all-cause) survival, $S(t)$, among the cancer patients divided by the expected survival, $S^*(t)$, the patients would have experienced had they not had cancer. The expected survival is typically obtained from nationwide population mortality rates (also called population life tables), stratified by age, sex, calendar year and possibly other covariates. The Human Mortality Database (HMD) [18, 19] contains population life tables for 37 countries, and the population life tables used for the studies in this thesis were all extracted from HMD. If the relative survival is lower than one, the cancer patients have a worse survival than a comparable group in the general population, and this is assumed to be due to the cancer under study. A relative survival of 1 suggests that the cancer patients have the same survival as a comparable group in the general population, but it does not mean that 100% of the cancer patients are still alive. In the relative survival setting the overall survival, as a function of time t since diagnosis, is written as

$$S(t) = S^*(t)R(t). \tag{6}$$

The main reason for using relative survival (instead of cause-specific survival where only deaths due to the cancer of interest are considered events and deaths due to other causes are censored) is that it does not rely on correct classification of cause of death. The cause of death can be poorly reported, especially among elderly patients [20], and even when the

reporting is good it can be difficult to determine if the death of a cancer patient is due to the cancer of interest or not (for example death from treatment complications) [21]. All studies included in this thesis use relative survival.

Relative survival will give an estimate of the net survival of interest if the cancer patients are exchangeable with the general population and there are no factors associated with both cancer and non-cancer mortality other than those factors that are both adjusted for in the analysis and in the population mortality file. If the cancer patients are not exchangeable with the general population the discrepancy between the observed and the expected survival can not be fully explained by the cancer and if the cancer and non-cancer mortality are not independent (given the factors adjusted for) the relative survival will not give an unbiased estimate of the net survival [22, 23, 24, 25].

The hazard analogue of relative survival is excess hazard, and it measures the mortality the patients experience in excess of what would be expected if they had not had cancer. The overall hazard, $h(t)$, among the patients is written as the sum of the expected hazard, $h^*(t)$, and the excess hazard, $\lambda(t)$, associated with the cancer

$$h(t) = h^*(t) + \lambda(t). \tag{7}$$

The overall as well as the relative survival can vary by covariates, $\boldsymbol{z}$, for example patient characteristics such as sex, age and calendar year of diagnosis, as well as tumour characteristics such as stage or grade. The expected mortality is allowed to vary by the stratification factors given in the population mortality rates, denoted with $\boldsymbol{z}'$, which are usually a subset of $\boldsymbol{z}$. Equation (6) and (7) are then extended as follows,

$$S(t;\boldsymbol{z}) = S^*(t;\boldsymbol{z}')R(t;\boldsymbol{z}), \text{and} \tag{8}$$

$$h(t;\boldsymbol{z}) = h^*(t;\boldsymbol{z}') + \lambda(t;\boldsymbol{z}). \tag{9}$$

Different regression models have been suggested for modelling the excess hazard, but most of the methods require follow-up time to be split into pre-specified time-bands or numerical integration [2, 26, 27, 28, 29]. Extensions of the Cox proportional hazards model have also been suggested for modelling of the excess hazard [30, 31], but unlike the standard Cox model the baseline is needed in the estimation. An alternative approach, which enables the modelling of time continuously and easily incorporates time-dependent effects, is the flexible parametric survival model [3, 4, 5]. The flexible parametric survival model has been extended to a relative survival setting [8]. The flexible parametric relative survival model is used in paper II-IV, and will be explained in the following sections.

2.6 The flexible parametric survival model

The flexible parametric survival model was first introduced by Royston and Parmar in 2001 [3, 4]. The model is fitted on the log cumulative hazard scale, using restricted cubic splines [6, 7] to estimate the baseline log cumulative hazard. Splines are piecewise polynomial functions (of order 3 for cubic splines) that are forced to have continuous first and second derivatives at the joining points, called knots. This ensures a smooth function. Restricted cubic splines are cubic splines forced to be linear beyond the boundary knots, to make the estimated function less influenced by sparse data in the tail of the distribution. Before extending the flexible parametric survival model to a relative survival framework in the next section, the general model is explained here.

The log cumulative hazard is modelled as a function of follow-up time, t, as

$$\ln(H(t)) = s(x; \boldsymbol{\gamma_0}) \tag{10}$$

where $x = \ln(t)$ and $s(x; \boldsymbol{\gamma_0})$ is a restricted cubic spline function. The latter is defined as

$$s(x; \boldsymbol{\gamma_0}) = \gamma_{00} + \gamma_{01} v_1(x) + \gamma_{02} v_2(x) + ... + \gamma_{0K-1} v_{K-1}(x), \tag{11}$$

where K is the number of knots and the p^{th} basis function is defined as

$$v_p(x) = \begin{cases} x, & \text{for } p = 1 \\ (x - k_p)_+^3 - \lambda_p (x - k_1)_+^3 - (1 - \lambda_p)(x - k_K)_+^3, & \text{for } p = 2, ..., K-1 \end{cases} \tag{12}$$

where $u_+ = u$ if $u > 0$ and $u_+ = 0$ if $u \leq 0$, k_1 is the position of the first knot, k_K the position of the last knot, k_p the position of the p^{th} knot, and $\lambda_p = \frac{k_K - k_p}{k_K - k_1}$.

Introducing covariates, $\boldsymbol{z}$, into equation (10) gives

$$\ln(H(t; \boldsymbol{z})) = s(x; \boldsymbol{\gamma_0}) + \boldsymbol{z}\boldsymbol{\beta}. \tag{13}$$

This is a proportional hazards model, so covariate effects, β, are interpreted in the same way as for models on the log hazard scale. If $s(x; \boldsymbol{\gamma_0})$ is a linear function of x, this is a Weibull model. So, the flexible parametric survival model is similar to a Weibull model but instead of modelling the log cumulative hazard as a linear function of log time it is modelled with the use of splines. Since a restricted cubic spline is used in the flexible parametric survival model, the log cumulative hazard is a linear function of log time after the last knot. Therefore, the flexible parametric survival model behaves like a Weibull distribution in the tail. Non-proportional hazards, i.e, time-dependent covariate effects, can be modelled by

including interactions between covariates and spline functions for log time. Modelling of the departure from a time-fixed effect usually does not require as many knots as modelling of the baseline cumulative hazard. Therefore a new set of spline parameters are introduced for each time-dependent effect, and separate knot positions can be chosen for each covariate, $\boldsymbol{z_i}$, with a time-dependent effect. This gives the model:

$$\ln(H(t;\boldsymbol{z})) = s(x;\boldsymbol{\gamma_0}) + \boldsymbol{z\beta} + \sum_{i=1}^{D} s(x;\boldsymbol{\gamma_i})\boldsymbol{z_i}, \tag{14}$$

where D is the number of time-dependent covariate effects and $s(x;\boldsymbol{\gamma_i})$ is the spline function for the i^{th} time-dependent effect.

Estimation is performed by maximum likelihood. The general log-likelihood for right-censored survival data can be expressed as:

$$\ln L = \sum_{i=1}^{N} d_i \ln[h(t_i;\boldsymbol{z_i})] + \ln[S(t_i;\boldsymbol{z_i})], \tag{15}$$

where N is the number of individuals, $d_i = 1$ for an event (e.g. death) and 0 for censored observations. Let $\eta = \ln(H(t;\boldsymbol{z}))$ where $\ln(H(t;\boldsymbol{z}))$ is expressed as in equation (13) or (14), then the survival function and the hazard function are expressed by their relationship with the log cumulative hazard function as

$$S(t;\boldsymbol{z}) = \exp(-\exp(\eta)), \text{and} \tag{16}$$

$$h(t;\boldsymbol{z}) = \frac{d}{dt}\exp(\eta). \tag{17}$$

The hazard function requires derivatives of the spline functions, and they can easily be obtained analytically [3]. Since both the survival and the hazard functions can be obtained analytically, numerical integration is not needed, which speeds up computation time. This is one of the reasons for fitting the flexible parametric model on the log cumulative hazard scale. Another reason is that the log cumulative hazard function is a relatively stable function, so it is easy to capture it's shape with the use of splines.

The number and location of the knots in the flexible parametric survival model has to be defined by the user, and in my experience 4-6 knots is usually sufficient to capture the shape of the log cumulative hazard. Flexible parametric survival models have been implemented in Stata (Statacorp, College Station, TX) [32] in the package `stpm2` [5]. In this implementation the default knot distribution is according to centiles of the log event times with the first knot at the first observed event time and the last knot at the last event time. If, for example, 6

knots are used the knots are placed at the 0th, 20th, 40th, 60th 80th and 100th percentile of event times. By placing the boundary knots at the first and last event times there is no linearity assumption within the range of the data. The user can instead of using the default positions explicitly position the knots, either by specifying other percentiles or by specifying time points (in years of follow-up) for the positions of the knots. Sensitivity to the knot distribution has been evaluated, and the flexible parametric survival model has been shown to be insensitive to both the number and location of the knots [33, 34].

2.7 The flexible parametric survival model for relative survival

The flexible parametric survival model has been extended to relative survival [8], and the extended model is used in studies II-IV. By integrating equation (9), the overall cumulative hazard, $H(t)$, is expressed as

$$H(t; \boldsymbol{z}) = H^*(t; \boldsymbol{z}') + \Lambda(t; \boldsymbol{z}), \tag{18}$$

where $H^*(t)$ is the cumulative expected hazard and $\Lambda(t)$ is the cumulative excess hazard. In a flexible parametric survival model for relative survival, $\Lambda(t; \boldsymbol{z})$ is modelled using splines in the same manner as described for $H(t; \boldsymbol{z})$ in section 2.6.

The log-likelihood for relative survival is expressed as

$$\ln L = \sum_{i=1}^{N} d_i \ln[h^*(t_i; \boldsymbol{z_i'}) + \lambda(t_i; \boldsymbol{z_i})] + \ln[S^*(t_i; \boldsymbol{z_i'})] + \ln[R(t_i; \boldsymbol{z_i})]. \tag{19}$$

Since $S^*(t_i; \boldsymbol{z}')$ is not dependent on any of the unknown model parameters it can be removed from the log likelihood,

$$\ln L = \sum_{i=1}^{N} d_i \ln[h^*(t_i; \boldsymbol{z_i'}) + \lambda(t_i; \boldsymbol{z_i})] + \ln[R(t_i; \boldsymbol{z_i})], \tag{20}$$

so the only background information needed is the expected rate at event times for those who die. The relative survival, $R(t; \boldsymbol{z})$, and excess hazard, $\lambda(t; \boldsymbol{z})$, can be obtained analytically as described in equations (16) and (17), where η now refers to the log cumulative excess hazard function, $\Lambda(t; \boldsymbol{z})$, as opposed to the log cumulative hazard function.

2.8 Cure models

For many types of cancer the mortality among the patients will eventually return to the same level as in the general population, i.e. $R(t)$ reaches a plateau (as seen in Figure 1) and

$\lambda(t)$ reaches zero at some time point after diagnosis. This point in time is called the cure point and the patients still alive at this point are considered "statistically cured". This is a population definition of cure and does not necessarily imply that all patients are medically cured. Statistical cure is not estimated on an individual basis, but gives an estimate of the proportion of patients that will not experience excess mortality due to the cancer, and can be of interest for measuring long-term survival in population-based cancer studies. The two most commonly used cure models in population-based studies are described in the following sections.

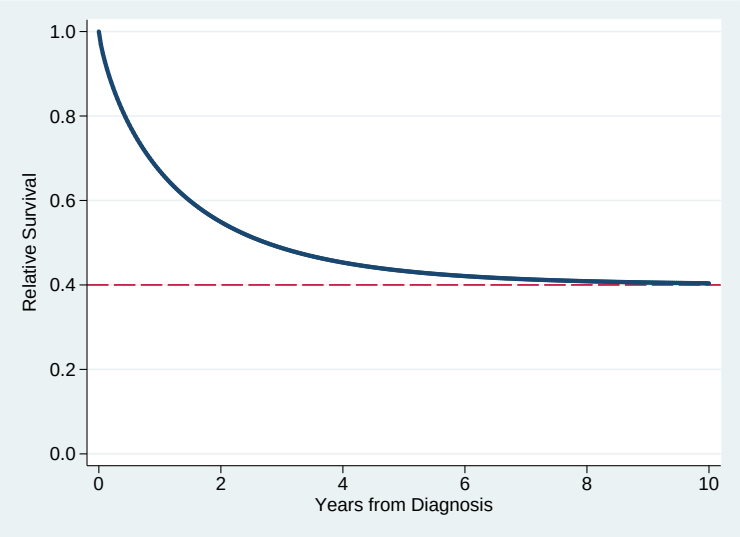

Figure 1 – Hypothetical relative survival curve where the estimated cure proportion is 0.4.

Mixture cure model

One of the most commonly used cure models in population-based cancer studies is the mixture cure model [35, 36, 37, 38, 39, 40, 41], and it is the model used in paper I. The model assumes that a proportion, π, are not at risk of experiencing the event (i.e. dying due to the diagnosed cancer) and are therefore considered cured. The other proportion, $1-\pi$, are the uncured and will, in the absence of censoring, experience the event. Their survival function associated with the disease will therefore tend to zero. In a relative survival setting the overall survival function for the mixture cure model is written as

$$S(t;\boldsymbol{z}) = S^*(t;\boldsymbol{z}')(\pi(\boldsymbol{z}) + (1-\pi(\boldsymbol{z}))S_u(t;\boldsymbol{z})), \tag{21}$$

where $S_u(t;\boldsymbol{z})$ is the cancer-specific survival function for the uncured, for which a parametric distribution is usually chosen, and a Weibull distribution is often used [37, 38, 42, 43, 44].

The covariates, $\boldsymbol{z}$, included for modelling the cure proportion, π, and the survival of uncured, $S_u(t)$, can be different. For example, the survival distribution of uncured can depend on only a subset of the covariates included to model the cure proportion, or the survival distribution of the uncured can be assumed to not vary by covariates at all. For simplicity, I will continue to refer to the covariates as $\boldsymbol{z}$ for both the cure proportion and the survival distribution of uncured, and in most cases it is recommended to include all covariates of interest in the modelling of both quantities. The dependence between the cure proportion, π, and the covariates can be modelled using different link functions;

1. the identity link, $\pi = \boldsymbol{z\beta}$. The covariate effects are in units of the cure proportion.

2. the logistic link, $\log(\pi/(1-\pi)) = \boldsymbol{z\beta}$. Covariate effects are expressed as log odds of cure.

3. the complementary log-log link, $\log(-\log(\pi)) = \boldsymbol{z\beta}$.

The overall (all-cause) hazard is the sum of the background mortality rate and the excess mortality rate associated with the cancer of interest

$$h(t;\boldsymbol{z}) = h^*(t;\boldsymbol{z}') + \frac{(1-\pi(\boldsymbol{z}))f_u(t;\boldsymbol{z})}{\pi(\boldsymbol{z}) + (1-\pi(\boldsymbol{z}))S_u(t;\boldsymbol{z})} \tag{22}$$

where $h^*(t;\boldsymbol{z}')$ is the expected mortality rate and $f_u(t;\boldsymbol{z})$ is the density function associated with $S_u(t;\boldsymbol{z})$. The log-likelihood (equation 15) becomes

$$\ln L = \sum_{i=1}^{N} d_i \ln\left(h^*(t_i;\boldsymbol{z_i'}) + \frac{(1-\pi(\boldsymbol{z_i}))f_u(t_i;\boldsymbol{z_i})}{\pi(\boldsymbol{z_i}) + (1-\pi(\boldsymbol{z_i}))S_u(t_i;\boldsymbol{z_i})}\right) + \ln(\pi(\boldsymbol{z_i}) + (1-\pi(\boldsymbol{z_i}))S_u(t_i;\boldsymbol{z_i})). \tag{23}$$

A mixture cure model with Weibull distribution and logistic link was used in paper I to study temporal trends in survival of patients with acute myeloid leukaemia.

Non-mixture cure model

Another cure model that has been proposed within a relative survival setting is the non-mixture cure model [39, 45], which estimates an asymptote for the survival function at the cure proportion. The non-mixture cure model was first introduced to model tumour recurrence [46, 47], but can also be used to model the statistical cure proportion. The survival function for the non-mixture model can be written as

$$S(t;\boldsymbol{z}) = S^*(t;\boldsymbol{z}')\pi(\boldsymbol{z})^{F_y(t;\boldsymbol{z})} = S^*(t;\boldsymbol{z}')\exp(\ln(\pi(\boldsymbol{z})) - \ln(\pi(\boldsymbol{z}))S_y(t;\boldsymbol{z})), \tag{24}$$

where $F_y(t;\boldsymbol{z})$ is a distribution function with $S_y(t;\boldsymbol{z})$ its corresponding survival function, and as for the mixture model, a Weibull distribution is often used. The overall hazard function is written as

$$h(t;\boldsymbol{z}) = h^*(t;\boldsymbol{z}') - \ln(\pi(\boldsymbol{z}))f_y(t;\boldsymbol{z}). \tag{25}$$

The non-mixture cure model can be expressed as a mixture cure model by rewriting equation (24) as

$$S(t;\boldsymbol{z}) = S^*(t;\boldsymbol{z}')\left(\pi(\boldsymbol{z}) + (1-\pi(\boldsymbol{z}))\left(\frac{\pi(\boldsymbol{z})^{F_y(t;\boldsymbol{z})} - \pi(\boldsymbol{z})}{1-\pi(\boldsymbol{z})}\right)\right) \tag{26}$$

and thus the survival distribution of the uncured can be estimated. The log-likelihood function becomes

$$\ln L = \sum_{i=1}^{N} d_i \ln(h^*(t_i;\boldsymbol{z_i'}) - \ln(\pi(\boldsymbol{z_i}))f_y(t_i;\boldsymbol{z_i})) + (\ln(\pi(\boldsymbol{z_i})) - \ln(\pi(\boldsymbol{z_i}))S_y(t_i;\boldsymbol{z_i})). \tag{27}$$

As in the previous section, the covariates included for modelling the cure proportion and the distribution function could be different, and the same link functions as described in the previous section can be used for modelling of the cure proportion. If the function $f_y(t)$ in equation (25) is not allowed to vary by covariates the non-mixture cure model is a proportional excess hazards model. This is a difference between the mixture and non-mixture cure models, since the former does not have a proportional excess hazards model for the whole group as a special case. If a complementary log-log link is used in a non-mixture cure model for which the function $f_y(t)$ do not vary by covariates (if proportional excess hazards can be assumed) the covariate effects are expressed as log excess hazard ratios. The non-mixture cure model can also be extended to use splines for modelling of the distribution function $F_y(t)$, which is done in paper II where the flexible parametric survival model is adapted to estimate cure.

Interpreting temporal trends using cure models

Cure models give estimates of both the cure proportion and the survival distribution of the uncured. By studying temporal trends in both these measures simultaneously, more can be understood about changes in survival than by for example only studying the 5-year RSR. A summary measure for the survival distribution of the uncured is usually presented together with the cure proportion, and the most often used summary measure is the median survival time of the uncured.

When studying temporal trends in cure and median survival time of uncured, the four scenarios illustrated in Figure 2 can often be used to facilitate the interpretation. These scenarios are (a) increased cure proportion and median survival time of uncured, (b) increased

cure proportion but decreasing median survival time of uncured, (c) increased median survival time of uncured and unchanged cure proportion or (d) increased cure proportion but unchanged median survival time of uncured. Even though these scenarios are a simplification of reality, they can still be useful when interpreting temporal trends in the two quantities.

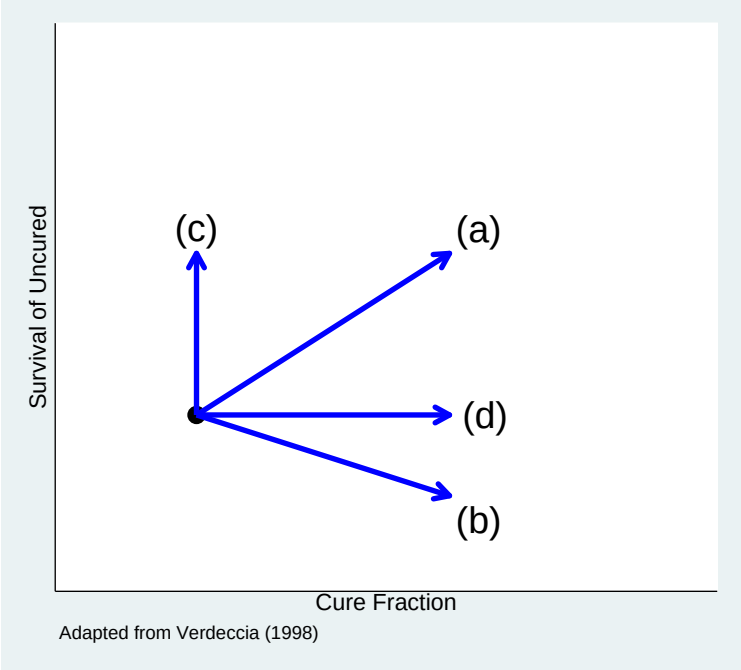

Figure 2 – Hypothetical changes in the cure fraction and median survival time of uncured between two calendar periods of diagnosis.

A general improvement of cancer patient survival might increase the chance of cure and also prolong life for those uncured, this is one possible explanation for observing scenario (a). When more patients are cured, those left in the uncured group are often those with worse prognosis. The cure proportion will then be increasing and the median survival time may decrease due to the selection of people to the 'cured group', this is seen as scenario (b). If treatment and management of cancer is improving even though the improvement does not lead to better possibilities of cure, scenario (c) can be observed. Another reason for observing scenario (c) could be lead time bias. Since the cure proportion does not depend on follow-up time it is not affected by lead time, and the cure proportion can therefore be used when lead time is a concern. This is why we chose to present the cure proportion as a complement to 5-year RSR in a study comparing survival between screen-detected and symptomatic cervical cancer in Sweden [48]. Scenario (d) could be seen if a new diagnostic procedure find patients with early stage disease that would never have been diagnosed if the procedure was not introduced. The new cases belong to the 'cured group', thus the cure proportion increase, but no change in the median survival time for the uncured is seen. In

reality, many things will simultaneously have an impact on the survival, and interpreting temporal trends always have to be done with care and all potential changes in diagnostic procedures should be taken into account.

2.9 Period analysis

To predict the survival for recently diagnosed patients, that have not yet been followed long enough to estimate their survival, a period approach to estimation has been suggested and proven empirically superior to the cohort approach [49, 50, 51, 52]. In a period analysis, only recently diagnosed patients contribute to the estimates of short term survival whereas patients diagnosed further in the past still contribute to estimates of long term survival, and therefore recent improvement in survival can be better captured. This set-up is made possible by pre-specifying a period window, and only person-time experienced within the period window contributes to the analysis. This is illustrated in figure 3, with a three year period window.

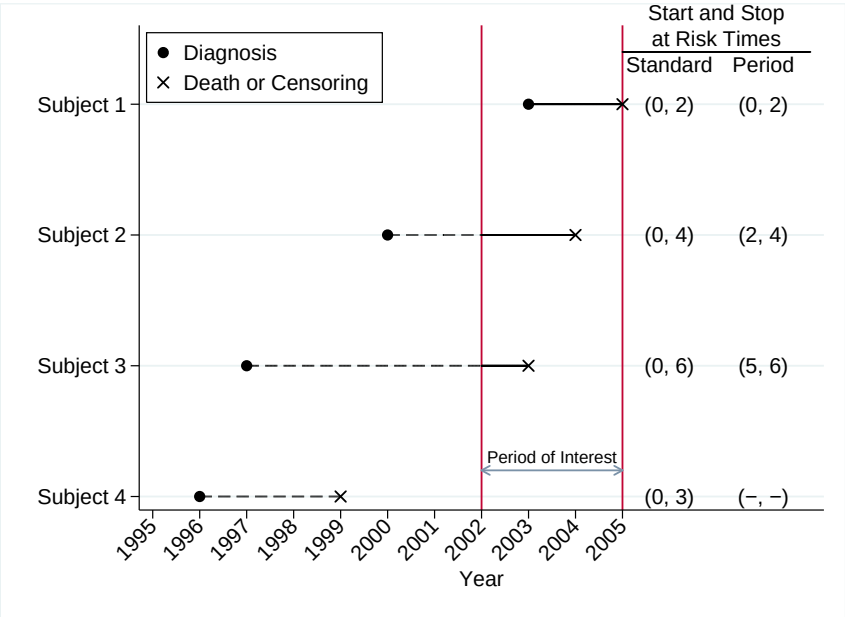

Figure 3 – Illustration of period analysis.

Subject number 1 is diagnosed within the period window and contributes follow-up time for the first 2 years after diagnosis. Subjects 2 and 3 are diagnosed prior to the period window and only contribute follow-up time to years 2-4 and 5-6, respectively. Subject number 4 is diagnosed prior to the period window, but also exits the study before the start of the period window and therefore does not contribute any follow-up time to the period analysis.

In survival analysis, a situation where all individuals are not followed from time 0 is

referred to as left truncation or delayed entry. Left truncation is easily incorporated into the likelihood by conditioning on being alive at the time point an individual enters the study. The log likelihood for a period analysis in a relative survival setting becomes

$$\ln L = \sum_{i=1}^{N} d_i \ln[h^*(t_i; \boldsymbol{z_i'}) + \lambda(t_i; \boldsymbol{z_i})] + \ln[R(t_i; \boldsymbol{z_i})] - \ln[R(l_i; \boldsymbol{z_i})], \tag{28}$$

where l_i is the time of entry for individual i. A period analysis was performed in paper IV to estimate the loss in expectation of life for recently diagnosed colon cancer patients in Sweden.

3 Aims of this thesis

The major objective of this PhD project was to develop and apply methods for presenting cancer survival statistics that are relevant for epidemiologists, clinicians and patients. More specifically:

- to extend the flexible parametric survival model to model the cure proportion among cancer patients.
- to extend the flexible parametric survival model to estimate the loss in expectation of life among cancer patients.
- to apply methods for population-based cancer patient survival to cancer registry data with the joint aim of evaluating the new methodology, presenting the new methodology to the broader research community, evaluating cancer care and gaining greater insights into temporal trends in cancer patient survival.
- to develop user friendly software to enable application of the methods.

4 Developed methods

4.1 Flexible parametric cure models

The cure models presented in section 2.8 require a parametric distribution to express either $S_u(t)$ or $F_y(t)$, and a Weibull distribution is often used. However, it has been shown that the Weibull distribution, and other distributions, are not always flexible enough to capture the shape of the underlying distribution [42, 53, 54]. More specifically, when the relative survival function drops substantially within a short time period after diagnosis (i.e. the excess hazard rate is initially very high) but later on reaches a plateau, the parametric distributions do not always give a good fit to the data. This type of scenario is often observed among older age groups, and is one reason why older patients have often been excluded from studies using cure models, for example in the study by Lambert *et al.* [42]. Another problem occur when the relative survival is high (survival of about 0.8 or above), for example among patients with localised disease, since the cure models often do not converge in these scenarios. A lot of alternative cure models have been suggested over the years, of which some are non- or semi-parametric (e.g, [55, 56, 57, 58]) or use a Bayesian approach (e.g, [59, 60]) or a spline-based approach (e.g, [61, 62, 63]). However, these approaches have not been developed in a relative survival setting. To overcome the shortcomings of standard relative survival cure models we adapted the flexible parametric relative survival model to estimate cure. This was published in paper II, together with an evaluation of how well the model performs in a situation where a mixture or non-mixture cure model with Weibull distribution did not give a good fit. A summary of paper II is given in section 6.2, but the method development is described here.

To estimate the cure proportion from the flexible parametric relative survival model we used the fact that when cure is reached the excess hazard rate is zero, and the cumulative excess hazard will therefore be constant after this point in time. By forcing the log cumulative excess hazard in the flexible parametric survival model to not only be linear after the last knot, but also to have zero slope, we impose a cure point and enable estimation of the cure proportion. This is done by manipulating the splines in the flexible parametric survival model. All spline functions except the linear, $v_1(x) = x$, are 0 before the first knot, so by imposing constraints on the parameter, γ_{01}, associated with this linear term it is possible to determine the slope of the spline before the first knot. To enforce cure, this has to be done after the last knot instead. By treating the knots in reversed order, all spline functions except the linear take the value 0 after the last knot instead of before the first. For lack of a better word, I refer to these new manipulated splines as "backwards" splines. The spline

basis functions, $v_p(x)$, in the backwards spline are defined as

$$v_p(x) = \begin{cases} x, & \text{for } p = 1 \\ (k_{K-p+1} - x)_+^3 - \lambda_p(k_K - x)_+^3 - (1 - \lambda_p)(k_1 - x)_+^3, & \text{for } p = 2, ..., K-1 \end{cases} \tag{29}$$

where $\lambda_p = \frac{k_{K-p+1} - k_1}{k_K - k_1}$. By using this backward spline in the flexible parametric survival model and constraining γ_{01} to 0, the log cumulative excess hazard, and therefore also the cumulative excess hazard, has 0 slope after the last knot and a cure point is imposed. The relative survival function for the flexible parametric survival model, with splines calculated backwards and with restriction for the linear spline parameter, from now on called the flexible parametric cure model, is defined as

$$R(t) = \exp(-\exp(\gamma_{00} + 0 \times v_1(x) + \gamma_{02} v_2(x) + ... + \gamma_{0K-1} v_{K-1}(x))), \tag{30}$$

which can be written as

$$R(t) = \pi^{\exp(\gamma_{02} v_2(x) + ... + \gamma_{0K-1} v_{K-1}(x))}, \tag{31}$$

where $\pi = \exp(-\exp(\gamma_{00}))$. When comparing to a non-mixture model we can see that the flexible parametric cure model is a special case of a non-mixture cure model with $\pi = \exp(-\exp(\gamma_{00}))$, and $F_y(t) = \exp(\gamma_{02} v_2(x) + ... + \gamma_{0K-1} v_{K-1}(x))$. $F_y(t)$ is a distribution function as long as the excess mortality is not negative, which is seen very rarely. As this is a non-mixture model, the flexible parametric cure model is a proportional excess hazards model, as long as no time-dependent effects are modelled.

When incorporating covariates,

$$R(t; \mathbf{z}) = \exp(-\exp(\gamma_{00} + \boldsymbol{z\beta}))^{(\exp(\gamma_{02} v_2(x) + ... + \gamma_{0K-1} v_{K-1}(x) + \sum_{i=1}^{D} s(x; \boldsymbol{\gamma_i}) z_i))} \tag{32}$$

we see that the constant parameters, γ_{00} and $\boldsymbol{\beta}$, are used to model the cure proportion and the time-dependent parameters are used to model the distribution function $F_y(t)$. The constraint of a zero effect for the linear spline term has to be incorporated for each time-dependent effect included in the model.

The constant parameter, γ_{00}, in equation (32) gives the log cumulative excess hazard at the time point of the last knot for the reference group and can be used to predict cure, since all other spline variables are 0 at the last knot. It is usually preferable to orthogonalise the spline variables, to avoid collinearity between the spline functions, which results in them not being zero from the point of the last knot. If orthogonalised splines are used, the same model

is being fitted, but the parameters take different values, and cure can then not be predicted by a direct transformation of the constant parameters. We therefore chose to center the orthogonalised spline variables around the value they take at the last knot, which enables direct predictions of cure from the constant parameters. Not only the cure proportion, but also the survival of uncured can be estimated from the flexible parametric cure model in the same way as for the non-mixture cure model, as seen in equation (26). The median survival time of uncured, or any other percentile, can be estimated using a Newton-Raphson algorithm.

We have adapted the Stata package for flexible parametric survival models (`stpm2`) [5], to include the flexible parametric cure model as a special case as well as post estimation commands to predict the cure proportion and the survival of the uncured. The updated Stata package is described in a paper that is not part of this thesis, but included as an appendix for completeness [64].

4.2 Life expectancy and loss in expectation of life

A useful summary measure for survival data is the life expectancy (LE), or mean survival time, which can be obtained by calculating the area under the survival curve. For cancer patients, the life expectancy from the date of the cancer diagnosis gives an estimate of the number of years a patient lives on average after they are diagnosed with cancer [65]. This is estimated by integrating the all-cause survival function

$$LE(\boldsymbol{z}) = \int_0^\infty S(u;\boldsymbol{z})du, \tag{33}$$

where $t = 0$ is the time of diagnosis. Note that it is the all-cause survival, $S(t)$ that is of interest, not the net survival. Integration should in theory be done up to ∞, but in practice a time point, t_{max}, is used where the survival function is assumed to have reached zero (i.e. all patients are assumed dead).

The loss in expectation of life (LEL) due to cancer is the difference between the expectation of life the patients would have had if they had not been diagnosed with cancer, estimated using population mortality rates for the general population, and the observed expectation of life among the cancer patients.

$$LEL(\boldsymbol{z}) = \int_0^{t_{max}} S^*(u;\boldsymbol{z}')du - \int_0^{t_{max}} S(u;\boldsymbol{z})du, \tag{34}$$

where $t = 0$ is the time of diagnosis and t_{max} is a point in time where both the expected and observed survival are assumed to have reached zero. Another measure is the proportion of

expected life lost (PELL), obtained by dividing the difference by the expectation of life,

$$PELL(\boldsymbol{z}) = \frac{\left(\int_0^{t_{max}} S^*(u;\boldsymbol{z}')du - \int_0^{t_{max}} S(u;\boldsymbol{z})du\right)}{\left(\int_0^{t_{max}} S^*(u;\boldsymbol{z}')du\right)}. \tag{35}$$

The loss in expectation of life, or the proportion of expected life lost, can be a useful measure for quantifying the cancer burden in the society or for an individual and for quantifying differences in survival between groups [65, 66].

The estimation of life expectancy generally requires extrapolation of the survival function beyond the available data, since the cohort of interest is in principle never followed up until the time point, t_{max}, when the survival function reaches zero. To estimate the loss in expectation of life, both the expected (general-population) survival and the observed all-cause survival (of the cancer patients) have to be extrapolated.

The expected survival can be extrapolated by making assumptions about the future mortality in the general population. The all-cause survival among the cancer patients can be extrapolated by assuming a parametric distribution, but it is difficult to find a statistical distribution that captures the underlying shape of the survival function, and even though a parametric distribution may fit well to the observed follow-up it may extrapolate poorly. Hakama and Hakulinen [67] suggested extrapolating the relative survival, and use the interrelationship between observed, expected and relative survival to obtain the extrapolated all-cause survival function. By breaking down the all-cause survival into two component parts, the expected survival and the relative survival, one can make assumptions for extrapolation of these functions independently. For most types of cancer the excess mortality is low after some years from diagnosis, so the expected mortality dominates for long-term follow-up, and the extrapolation therefore mostly depends on the extrapolation of the expected survival. Hakama and Hakulinen estimated loss in expectation of life using grouped data (life tables of relative survival) and suggested to extrapolate the relative survival by assuming statistical cure or by assuming that the excess mortality had stabilised to a constant. Other approaches that use background mortality have been suggested, of which some assume that background mortality alone can be used beyond the available follow-up [65, 68], and other extrapolate based on a linear regression model on the logit of relative survival or quality-adjusted survival [69, 70]. Nelson *et. al* use the observed data to estimate rates as a function of age and use that for extrapolation [71] beyond the available follow-up. Others have used parametric distributions to extrapolate the survival function [72, 73, 74, 75].

A similar measure, more commonly reported, is the years of life lost. Years of life lost is estimated as the difference between the age at death and the life expectancy of each patient

or a pre-specified cut-off age [66, 76, 77, 78]. However, this approach relies on accurate cause of death information to identify individuals in the population who died due to cancer. Since the approach only includes patients who have died, and specifically only those who die of their cancer, irrespective of when they were diagnosed, it can not be used for a specific cohort of patients. In addition, if a cut-off is used, any differences between groups that occur after the cut-off age are ignored. Life expectancy and the loss in expectation of life do not suffer from the same limitations.

In paper III, we demonstrate how the loss in expectation of life can be estimated by the use of a flexible parametric survival model. A flexible parametric survival model for the cumulative all-cause mortality, $H(t)$, can be used for extrapolating beyond the available follow-up, but the fit is not always good. Instead, similarly to Hakama and Hakulinen, we suggest to model the cumulative excess mortality part using a flexible parametric survival model. Three possible approaches for the extrapolation from a flexible parametric survival model for relative survival are:

1. Linear trend. Since restricted cubic splines are linear beyond the boundary knots, the flexible parametric survival model gives a linear log cumulative excess hazard beyond the last knot, or equivalently behaves like a Weibull distributed excess hazard in the tail. The linear trend is mainly based on the observed trend towards the end of follow-up.

2. Cure. Assuming statistical cure beyond the last boundary knot, by the use of a flexible parametric cure model presented in section 4.1, thus after the cure point we can use the expected survival alone when extrapolating.

3. Constant excess hazard. Assuming a constant excess hazard beyond the last boundary knot. To incorporate an assumption of constant excess hazard a similar approach as for the flexible parametric cure model is used. The spline variables are again calculated "backwards" and the parameter for the linear spline variable, γ_{01}, is set to 1. This gives an excess hazard that behaves like an exponential distribution in the tail. For this approach the smoothness of the splines can be relaxed slightly, which is described in section 4.3.

Based on the model the extrapolated relative survival can be estimated. When the full (extrapolated) relative survival function has been estimated the full (extrapolated) all-cause survival function can be estimated by multiplying the relative survival by the (extrapolated) expected survival function. The loss in expectation of life is then estimated as

$$LEL(\boldsymbol{z}) = \int_0^{t_{max}} S^*(u; \boldsymbol{z}')du - \int_0^{t_{max}} R(u; \boldsymbol{z})S^*(u; \boldsymbol{z}')du. \tag{36}$$

The post estimation command for the Stata package `stpm2` (package for flexible parametric survival models in Stata) has been extended to enable estimation of the life expectancy and loss in expectation of life. In the Stata package, the integrals are obtained numerically using the Gaussian quadrature rule [79] and the variance of the loss in expectation of life is obtained using the delta method [80].

4.3 Making splines more flexible

Restricted cubic splines are by design very smooth, and that is partly why they are so popular. In most applications of flexible parametric survival models the smoothness of the restricted cubic splines is an advantage. However, when one wants to change the behaviour of the splines in the tail of the distribution, as in the flexible parametric cure model or to impose a constant excess hazard after a certain point (which has been suggested for estimation of loss in expectation of life), the smoothness of the splines might introduce incorrect shapes of the cumulative hazard even prior to the point where the restriction is imposed. This primarily happens when the imposed constraint forces the log cumulative (excess) hazard function to change from a concave to a convex function or vice versa. Since the log cumulative excess hazard function is usually a concave function the problem will arise when a constant excess hazard is imposed after the last knot, but not so often when imposing a cure point. This is illustrated graphically in figure 4, where the log cumulative excess hazard function from a model without constraint, a model with constant excess hazard after 8 years post diagnosis and a model with a cure point at 8 years post diagnosis is plotted against log time, together with their corresponding cumulative excess hazard, excess hazard and relative survival functions. For all three models, the last knot was positioned at 8 years after diagnosis, and this point is shown in the graphs as a vertical line. In the flexible parametric relative survival model the log cumulative excess hazard is modelled as a function of log time using splines, which is why we chose to show the log cumulative excess hazard against log time even though it would not usually be the scale of interest.

To overcome this potential problem, I relaxed the conditions of the restricted cubic spline by not requiring a continuous first and second derivative at the last knot. To construct a restricted cubic spline without continuous first and second derivative at the last knot I started with constructing a restricted cubic spline without continuous first and second derivative at the first knot. Firstly, a general cubic spline with K knots is specified as

$$s(x;\boldsymbol{\gamma}) = \gamma_{01}x + \gamma_{02}x^2 + \gamma_{03}x^3 + \sum_{p=2}^{K-1}\left(\gamma_p(x-k_p)_+^3\right) + \gamma_1(x-k_1)_+^3 + \gamma_K(x-k_K)_+^3, \quad (37)$$

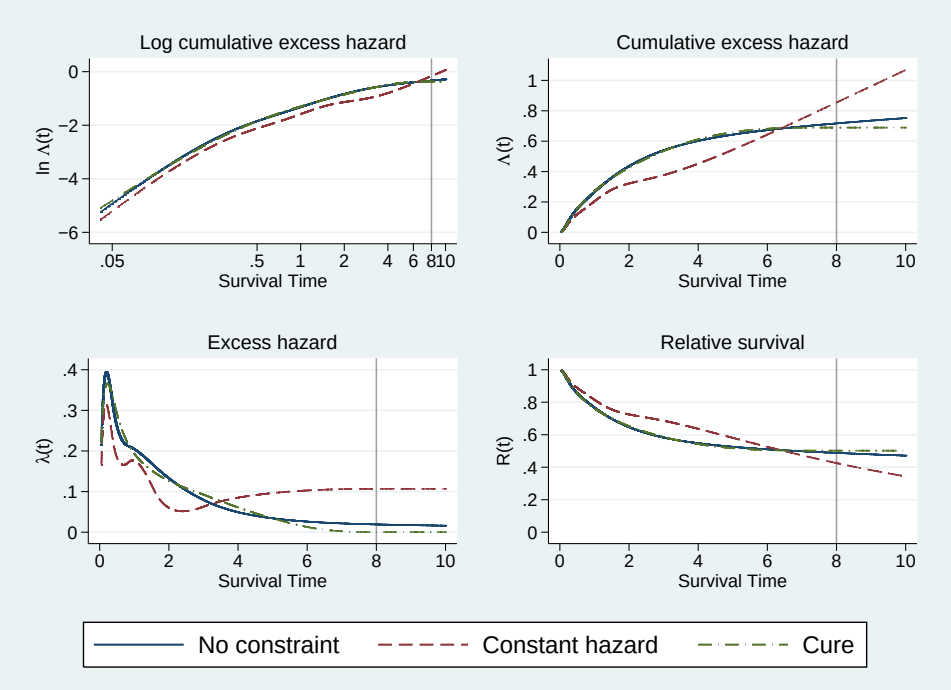

Figure 4 – Log cumulative excess hazard, cumulative excess hazard, excess hazard and relative survival functions from a standard flexible parametric survival model, a model with constant excess hazard after 8 years and a model with a cure point at 8 years.

where k_p refers to the p^{th} knot [6]. A cubic spline (with K knots) without the constraint of continuous second derivative at the first knot is written as

$$s(x;\boldsymbol{\gamma}) = \gamma_{01}x+\gamma_{02}x^2+\gamma_{03}x^3+\sum_{p=2}^{K-1}\left(\gamma_p(x-k_p)_+^3\right)+\gamma_1(x-k_1)_+^3+\gamma_K(x-k_K)_+^3+\gamma_{extra}(x-k_1)_+^2. \tag{38}$$

So we now have one extra parameter γ_{extra} and the extra spline variable corresponding to this parameter $(x-k_1)^2$. A cubic spline (with K knots) without constraint of continuous first or second derivative at the first knot is written as

$$\begin{aligned} s(x;\boldsymbol{\gamma}) =& \gamma_{01}x+\gamma_{02}x^2+\gamma_{03}x^3+\sum_{p=2}^{K-1}\left(\gamma_p(x-k_p)_+^3\right)+\gamma_1(x-k_1)_+^3+\gamma_K(x-k_K)_+^3+ \\ & \gamma_{extra1}(x-k_1)_+^2+\gamma_{extra2}(x-k_1)_+. \end{aligned} \tag{39}$$

So we now have two extra parameters γ_{extra1} and γ_{extra2} and the extra spline variables corresponding to these parameters are $(x-k_1)^2$ and $(x-k_1)$. Restricted cubic splines are forced to be linear before the first knot and after the last knot. To impose linearity before the first knot we need to set $\gamma_{02}=\gamma_{03}=0$. To impose linearity after the last knot we use

the fact that $s''(x) = s'''(x) = 0$ for linear functions. So for $x > k_K$

$$s'(x;\boldsymbol{\gamma}) = \gamma_{01} + 3\sum_{p=2}^{K-1}\big(\gamma_p(x-k_p)^2\big) + 3\gamma_1(x-k_1)^2 + 3\gamma_K(x-k_K)^2 + 2\gamma_{extra1}(x-k_1) + \gamma_{extra2} \tag{40}$$

$$s''(x;\boldsymbol{\gamma}) = 6\sum_{p=2}^{K-1}\big(\gamma_p(x-k_p)\big) + 6\gamma_1(x-k_1) + 6\gamma_K(x-k_K)^2 + 2\gamma_{extra1} \tag{41}$$

$$s'''(x;\boldsymbol{\gamma}) = 6\sum_{p=2}^{K-1}\gamma_p + 6\gamma_1 + 6\gamma_K \tag{42}$$

Setting $s''(x) = s'''(x) = 0$ gives:

$$s'''(x;\boldsymbol{\gamma}) = 0 \implies \gamma_K = -\sum_{p=2}^{K-1}\gamma_p - \gamma_1 \tag{43}$$

$$s''(x;\boldsymbol{\gamma}) = 0 \implies \gamma_1(x-k_1) = -\sum_{p=2}^{K-1}\big(\gamma_p(x-k_p)\big) - \gamma_K(x-k_K) - \gamma_{extra1}/3 \tag{44}$$

$$\begin{aligned} &\text{substituting } \gamma_K \implies \\ \gamma_1(x-k_1) &= -\sum_{p=2}^{K-1}\big(\gamma_p(x-k_p)\big) + \sum_{p=2}^{K-1}\big(\gamma_p(x-k_K)\big) + \gamma_1(x-k_K) - \gamma_{extra1}/3 \\ &\implies \gamma_1 = -\sum_{p=2}^{K} -1\gamma_p\lambda_p - c\gamma_{extra1} \\ &\implies \gamma_K = -\sum_{p=2}^{K} -1\big(\gamma_p(1-\lambda_p)\big) + c\gamma_{extra1} \end{aligned} \tag{45}$$

where $\lambda_p = (k_K - k_p)/(k_K - k_1)$ and $c = 1/(3(k_K - k_1))$. A restricted cubic spline with relaxed smoothness at the first knot is therefore expressed as

$$\begin{aligned} s(x;\boldsymbol{\gamma}) =& \gamma_{01}x + \sum_{p=2}^{K-1}\Big(\gamma_p\big((x-k_p)_+^3 - \lambda_p(x-k_1)_+^3 - (1-\lambda_p)(x-k_K)_+^3\big)\Big) + \\ & \gamma_{extra1}\big((x-k_1)_+^2 - c(x-k_1)_+^3 + c(x-k_K)_+^3\big) + \gamma_{extra2}(x-k_1)_+, \end{aligned} \tag{46}$$

where $\lambda_p = (k_K - k_p)/(k_K - k_1)$, $c = 1/(3(k_K - k_1))$ and k_p refers to the p^{th} knot.

To construct a cubic spline with relaxed smoothness at the last knot, we want to calculate equation (46) backwards, i.e. treating the knots in reversed order, so that all variables except

the linear term are 0 after the last knot instead of before the first knot. This is done in the same way as for standard restricted cubic splines, as explained in section 4.1, but with the extra variables also calculated in the reversed way. Backwards restricted cubic splines (with K knots) without the constraints of continuous first or second derivative at the last knot is written as

$$\begin{aligned} s(x;\boldsymbol{\gamma}) =& \gamma_{01}x + \sum_{p=2}^{K-1} \left(\gamma_p \left((k_{K-p+1}-x)_+^3 - \lambda_p(k_K-x)_+^3 - (1-\lambda_p)(k_1-x)_+^3\right)\right) + \\ & \gamma_{extra1}\left((k_K-x)_+^2 - c(k_K-x)_+^3 + c(k_1-x)_+^3\right) + \gamma_{extra2}(k_K-x)_+, \end{aligned} \tag{47}$$

where $\lambda_p = (k_{K-p+1}-k_1)/(k_K-k_1)$ and $c = 1/(3(k_K-k_1))$. A difference between the new, more flexible, restricted cubic splines and standard restricted cubic splines is that the forward and backwards spline will not give the same model with the new splines, since the relaxed constraints are in different places.

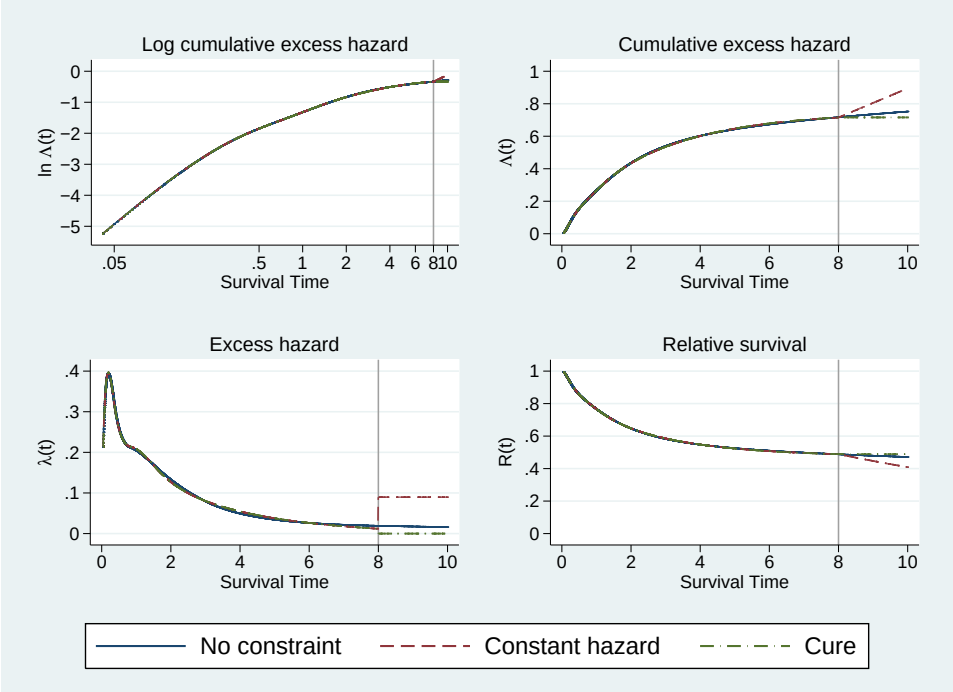

Figure 5 – Log cumulative excess hazard, cumulative excess hazard, excess hazard and relative survival functions from a standard flexible parametric survival model, a model with constant excess hazard after 8 years and a model with a cure point at 8 years. The two latter are modelled with splines allowed to have non-continuous first and second derivative at the last knot.

Results using the new more flexible spline is seen in figure 5, where the models with a constant excess hazard and cure are refitted using the more flexible splines. The three models now agree very well up until the position of the last knot. The hazard function is the

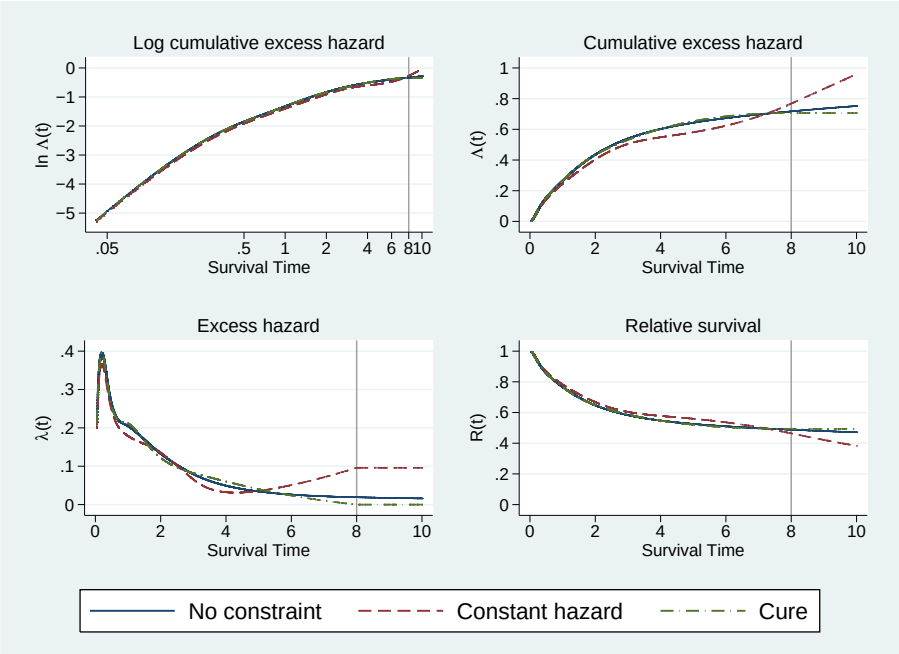

Figure 6 – Log cumulative excess hazard, cumulative excess hazard, excess hazard and relative survival functions from a standard flexible parametric survival model, a model with constant excess hazard after 8 years and a model with a cure point at 8 years. The two latter are modelled with splines allowed to have non-continuous second derivative at the last knot.

derivative of the cumulative hazard function and, due to the relaxation of the smoothness, the derivative of the cumulative hazard function is not defined at the position of the last knot, and therefore the hazard function can jump in the way seen in figure 5. In this example a constant excess hazard is not a reasonable assumption, which makes this a good example for illustrating the potential problem with imposing constraints after the last knot, and the jump would be smaller if a constant excess hazard was more reasonable. However, we have not found a dataset where it seems plausible to assume a constant excess hazard, and where the new spline fit well. Therefore, I would not recommend the use of these splines for this purpose without thorough sensitivity analyses. Even so, the more flexible splines were used in paper III to fit models that assumed a constant excess hazard after the last knot, since they at least do not impose implausible shapes of the cumulative excess hazard function prior to the last knot.

Instead of relaxing both the first and second derivative at the last knot, splines with only non-continuous second derivative at the last knots can be used. The exact description will not be given here, but the construction of these splines is similar to the description above. These splines could potentially be advantageous to use within the flexible parametric cure model, but they are not smooth enough to overcome the problem of an inflection point when assuming a constant excess hazard, as seen in figure 6.

5 Materials

5.1 The Swedish Cancer Registry

The Swedish Cancer Registry was established in 1958, and all health care providers and pathologists are obliged by law to notify the registry about all new cases of cancer. Since the 1980's the registration and coding is done at six regional registries located in each of the six Swedish health care regions. Information is collected by the regional registries on patient characteristics such as age and sex, as well as tumour characteristics such as stage and histopathology. The cancer registries also obtain follow-up information on date and cause of death as well as date of emigrations. This is done by linking the data to other registries by the use of the ten-digit ID number allocated to all residents in Sweden. For most diagnosed cancers the regional registry will receive two independent reports, one from the pathologist and one from the clinician. The regional cancer registries annually supply data to the national Swedish Cancer Registry, and the completeness of the register is above 96% [81]. The national cancer register include less detailed information than the regional registries, and stage at diagnosis was not reported to the national register until 2004. Data from the Swedish cancer register was used in studies I, III and IV. Paper I includes diagnoses of acute myeloid leukaemia (AML), paper III diagnoses of breast cancer, colon cancer, malignant melanoma as well as bladder cancer and paper IV includes diagnoses of colon cancer. More detailed information on the data used in each paper is given in the brief summary of each paper in section 6.

5.2 The Finnish Cancer Registry

The Finnish Cancer Registry started in 1953, and cancer reporting has been compulsory since 1961. The register includes data about the patient as well as information about the tumour, and the completeness of registration is over 99% [82]. The underlying cause of death is recorded for all cases, using death certificate information from Statistics Finland by linkage through the ten-digit ID number allocated to all residents. Death certificate information from Statistics Finland is also used to add cancers notified on death certificates that have not previously been recorded in the cancer register, i.e. death certificate only (DCO) cases. This is a difference between the Finnish and Swedish cancer registries since the latter does not include DCO cases. The Finnish Cancer Registry is also an active research institute that performs statistical and epidemiological cancer research. Data on colon cancer in the Finnish cancer register was used in paper II, more detailed information is given in section 6.

6 Summary of papers

6.1 Paper I

Background

Acute myeloid leukaemia (AML) is an aggressive disease and rapidly fatal if left untreated. Curative treatment is now available which has led to improved patient survival. Large age-dependent differences in AML patient survival and greater improvement over time for younger patients have been shown in Sweden [83]. In that study AML patient survival was presented as 1- and 5-year RSR for different age groups and calendar periods. In this study we reanalysed the data using cure models to gain further insights into changes observed in AML patient survival over time and how these changes vary between different age groups.

Material

We identified all AML diagnoses, International Classification of Diseases revision 7 (ICD-7) codes 2050, 2059, 2060, 2069 and recorded as malignant, in the Swedish Cancer Registry between 1973–2001. If the same individual had more than one AML diagnosed within this time period, only the first was included. Incidental autopsy findings, in total 330 cases, were excluded. Patients 18 years or younger and 81 years or older at the time of diagnosis were excluded, which led to the exclusion of 1 604 patients. AML is not common among children and therefore we chose to restrict to adult patients, and the reason to exclude patients older than 80 was that the cure models tend to be less reliable for older age groups, as described in section 4.1. After also excluding 35 patients with reported date of death or emigration before the date of diagnosis, the study population consisted of 6 439 patients. Patients were followed until death or censored 10 years after diagnosis, at date of emigration or December 31 2006, whichever occurred first.

Methods

A mixture cure model (described in section 2.8) was used with a Weibull distribution for the survival function of the uncured and a logistic link for the cure proportion. The variables of interest were age at diagnosis and year of diagnosis. We chose to include age as a categorical variable, with the same categorisation as in the previous study [83] (19-40, 41-60, 61-70, 80 years and above). In the previous study calendar period was also included as a categorical variable, but in this study we chose to model calendar year as a continuous variable using restricted cubic splines (with four knots) to gain further insights into the temporal trends in AML patient survival. The cure proportion as well as both Weibull parameters were allowed

to vary by both age group and calendar year and an interaction between age group and calendar year was included. The analysis was carried out using the command `strsmix` [53] in Stata (Statacorp, College Station, TX).

Results

Both the cure proportion and the median survival time of uncured were low at the start of the study period (figure 7) for all age groups. But a dramatic improvement is seen for the youngest age group, with the cure proportion increasing from 4% (95% CI 2-10) in 1975 to 68% (95% CI 56-77) in 2000 and the median survival time of the uncured increasing from 0.43 years (95% CI 0.32-0.58) in 1975 to 1.08 years (95% CI 0.90-1.31) in 1990 and then decreased to 0.74 years (95% CI 0.43-1.26) in 2000. For the second age group (age 41-60 years at diagnosis) the cure proportion started to increase during the mid 1980s and was 32% (95% CI 25-39) in 2000 and the median survival time of the uncured increased to about 1 year in the mid 1990s and then decreased slightly. There were small improvements seen in the cure proportion of the third age group (61-70 years) and the cure proportion in 2000 was 8% (95% CI 3-21), while the median survival time of the uncured increased to about 0.7 years. Only small improvement in the median survival time of the uncured was seen for those aged 71-80 years at diagnosis, and no improvement in the cure proportion.

Sensitivity analysis

Whenever cure models are used, the assumption of an existing cure point should always be assessed. This was also performed within this study, but the results were not presented in detail in the paper. A mixture cure model was fitted to each age group, separately for each of the calendar periods 1973-1981, 1982-1990, 1991-2001. The relative survival estimated from the cure models was compared to empirical life table estimates of relative survival, to compare the fit as well as to evaluate the cure assumption. The results can be seen in figure 8. The empirical relative survival estimates reach a plateau, indicating statistical cure, and the relative survival estimated from the cure models closely follow the life table estimates of relative survival.

Conclusion

We saw large improvement in the cure proportion among younger ages, whereas improvement in the oldest ages is mainly within the survival of the uncured. The improvement in the cure proportion was seen earlier in the younger age groups than in older patients which probably can be explained by the fact that treatment with curative intent started earlier in younger

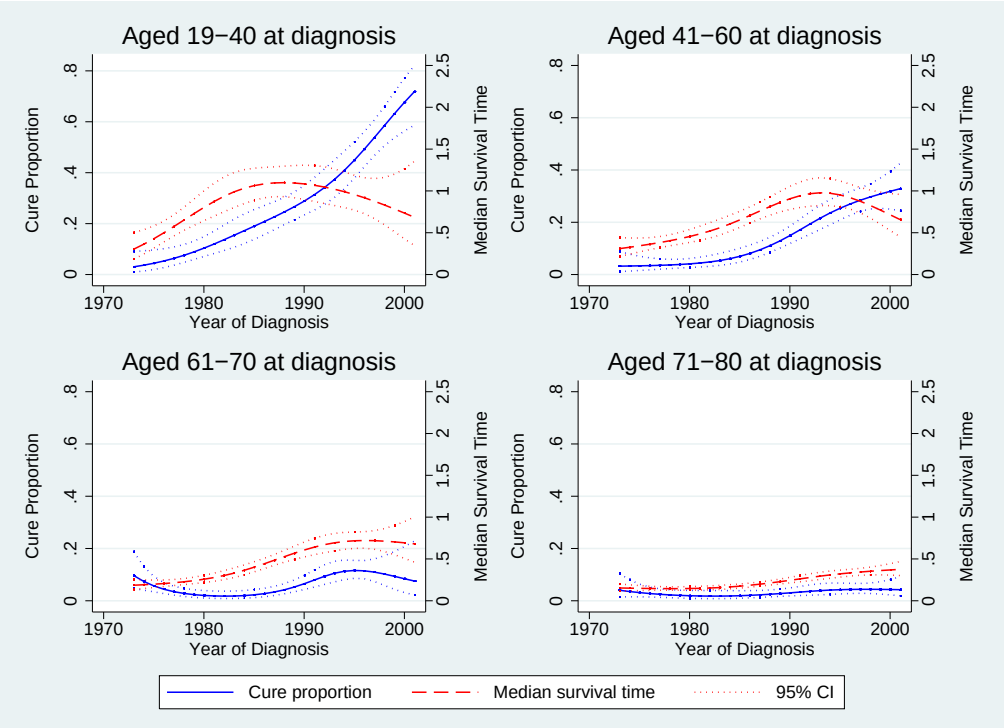

Figure 7 – Temporal trends in cure proportion and median survival time (in years) of uncured for patients diagnosed with AML in Sweden 1973-2001, presented by age group and with 95% confidence intervals.

patients and was later on given to an increasing proportion of elderly patients. The median survival time of the uncured decreased or was stable during the last ten years in all age groups. This could be due to better selection of curable patients leaving poor risk patients to mainly palliative care. Another explanation might be the highly toxic treatment, a patient treated with curative intent has a good chance of being cured, but if the treatment is unsuccessful it might shorten his or her life. Taken together, the information achieved by the use of a cure model improved the understanding of temporal trends in AML patient survival.

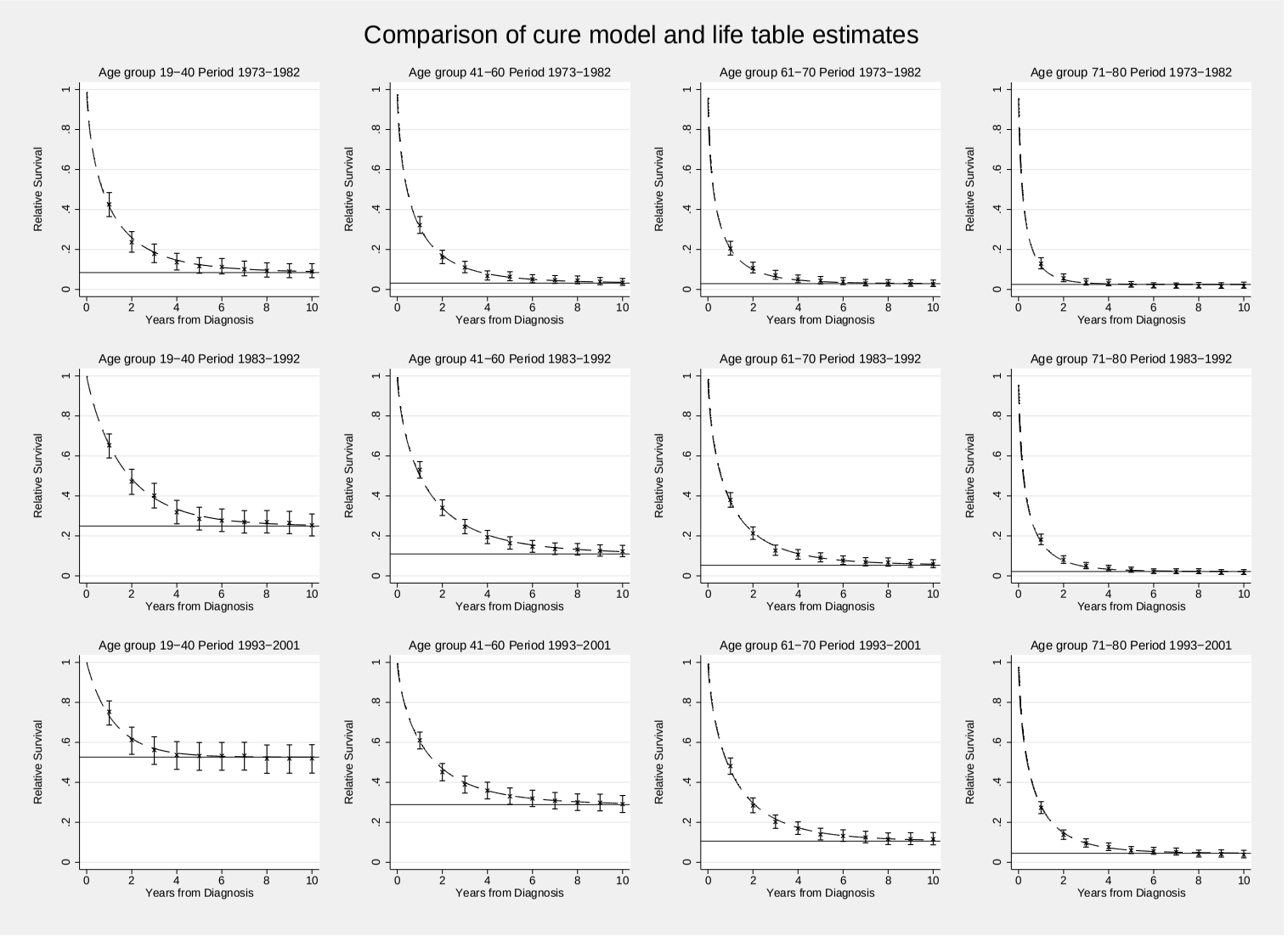

Figure 8 – Predicted relative survival (dashed lines) and cure proportions (solid lines) from mixture models compared to life table estimates (x) of relative survival, with 95% CI, for patients diagnosed with AML in Sweden 1973-2001.

6.2 Paper II

Background

Cure models can be a valuable tool to gain insights in cancer patient survival, how it changes over time or differs between groups. In population-based studies parametric cure models, as described in section 2.8, are popular since they can easily be fitted within a relative survival framework. A disadvantage of standard parametric cure models is that the distributional assumption does not always give a good fit to the data. In this study we extended the flexible parametric survival model to enable cure modelling with the use of splines. The new method is described in section 4.1, and the evaluation of the method is presented in this section.

Material

In a study by Lambert *et al.* [42] a mixture cure model with a Weibull distribution was used to study temporal trends in cancer of the colon and rectum in Finland. In that study patients aged 80 and older at diagnosis where excluded due to a poor fit of the cure model. In this study we reexamined the data on colon cancer, including patients aged 80 and older, using the flexible parametric cure model to compare the fit of the new model with the mixture and non-mixture cure model.

Methods

In the study by Lambert *et al.* a mixture cure model with Weibull distribution was used, including the variables year of diagnosis (modelled continuously with splines) and age at diagnosis (in categories less than 50, 50-59, 60-69 and 70-79) as well as an interaction between age group and year. All three model parameters, the cure proportion and the two Weibull parameters, were allowed to vary by covariates. In this study, since the flexible parametric cure model is a special case of a non-mixture cure model, we compared the flexible parametric cure model to a non-mixture cure model instead of a mixture cure model, but the two parametric models give very similar results. A Weibull non-mixture cure model and a flexible parametric cure model were fitted including the same covariates as in the previous study, but also including the age group 80 and older. Estimated cure proportions and median survival times of uncured were compared. The fit of the cure models were also compared to empirical life table estimates of relative survival, separately for each age group and calendar periods 1953-1964, 1965-1974, 1975-1984, 1985-1994 and 1995-2003. To investigate how the flexible parametric cure model performs in a situation where survival is high a subset analysis, restricted to localised cancer only, was performed. The fit of the non-mixture

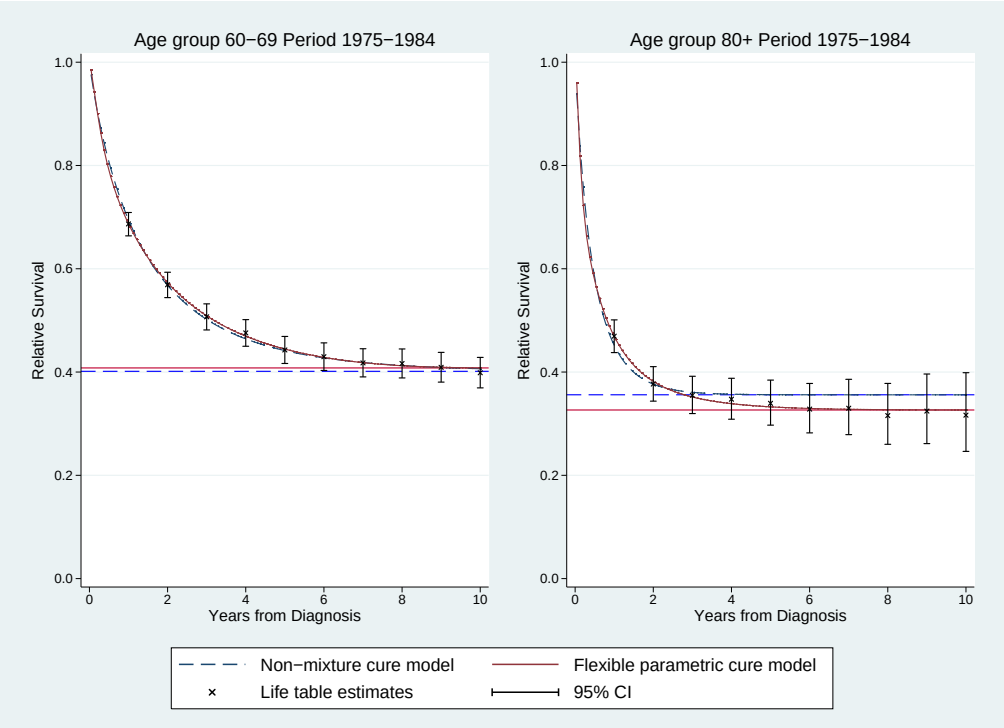

Figure 9 – Predicted survival and cure proportions (%) from non-mixture models and flexible parametric cure models, compared to life table estimates of relative survival.

and flexible parametric cure model were again compared to empirical life table estimates of relative survival, separately for each age group and calendar period.

Results

Figure 9 shows the predicted relative survival and cure proportions from non-mixture cure models and flexible parametric cure models fitted to the age groups 60-69 and 80 and older during calendar period 1975-1984, together with empirical life table estimates of relative survival. The predicted survival and cure proportions from the two cure models agree well for age group 60-69, and also correspond well with the life table estimates. For the oldest age group on the other hand, the two cure models do not agree as well, and the flexible parametric cure model follows the empirical life table estimates much more closely, giving a much better fit to the data. The difference between the two cure models is also seen in figure 10, where the cure proportion (together with 95 % CI), from models where calendar year is modelled using splines, is plotted for the same two age groups. The cure proportions are similar for the age group 60-69, but for the oldest age group where the parametric cure model has been shown to overestimate cure, the flexible parametric cure model gives lower estimates of cure. When restricting the data to localised cancer only, the Weibull non-mixture model did not converge for all combinations of age and calendar period. However, the flexible

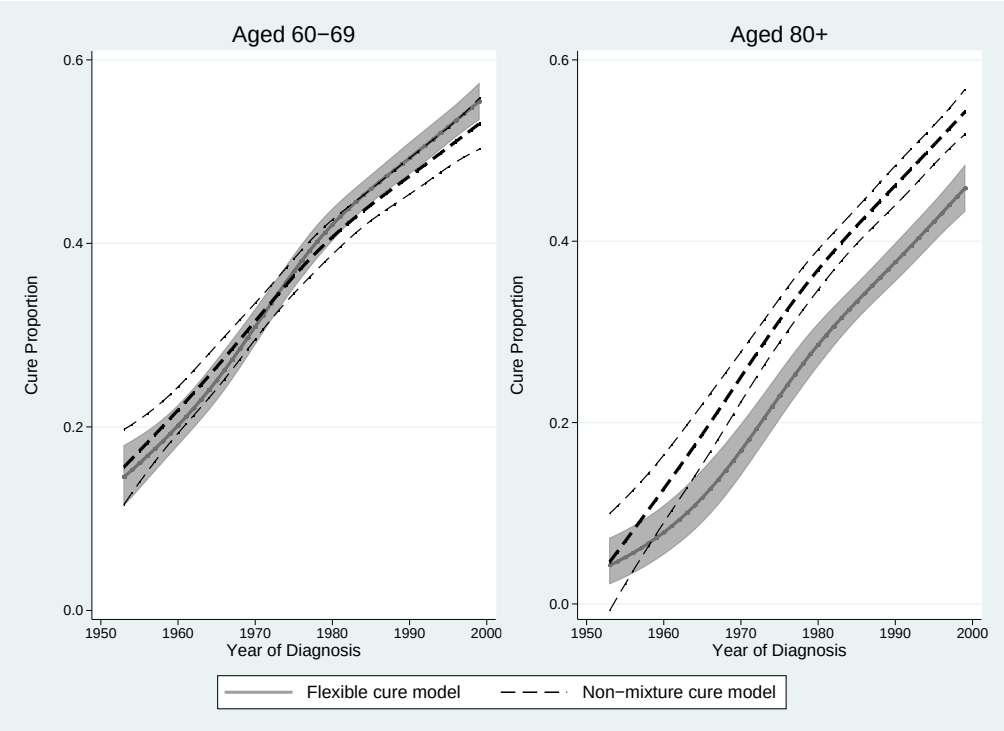

Figure 10 – Predicted cure proportions (%) with 95% CI from a Weibull non-mixture cure model and a flexible parametric cure model.

parametric cure model gave sensible estimates (when compared to life table estimates of relative survival) in all scenarios, and always converged.

Conclusion

The flexible parametric cure model enables estimation of cure when parametric cure models are not flexible enough, which enables for example, inclusion of older age groups. The `stpm2` Stata package for estimation of flexible parametric survival models [5] has been updated to enable the use of the proposed method [64].

6.3 Paper III

Background

Life expectancy and loss in expectation of life can be used to address a wide range of research questions of public health interest pertaining to the prognosis of cancer patients, but due to the need for extrapolation of the survival curve they are rarely reported. In this study we showed that the flexible parametric survival model can reliably extrapolate relative survival. By using the relationship between all-cause, relative and expected survival, the extrapolated all-cause survival can also be reliably estimated. This approach can then be used to estimate life expectancy and loss in expectation of life for cancer patients. The theory is described in section 4.2, and the evaluation of the extrapolation is presented in this section.

Material and methods

Data on colon cancer, breast cancer, malignant melanoma and bladder cancer diagnosed in Sweden 1961–1970 were used to evaluate how well the flexible parametric survival model extrapolates survival. The observed mean survival time based on Kaplan-Meier estimates (K-M) [84] of the available 40 years of follow-up was estimated and compared to mean survival times estimated from extrapolated survival functions where follow-up was restricted to 10 years for the analysis and extrapolated to 40 years. This was done separately for each cancer site and four age groups (50-59, 60-69, 70-79 and 80 years and above). Extrapolation was performed based on 5 different models:

1. extrapolating from a flexible parametric survival model for all-cause survival,
2. extrapolating from a flexible parametric survival model for relative survival,
3. extrapolating from a flexible parametric cure model for relative survival,
4. extrapolating from a flexible parametric survival model assuming constant excess hazard from the point of the last knot,
5. extrapolating from a model assuming that the excess hazard follows a Weibull distribution.

For more information about the different extrapolation approaches see section 4.2.

Results

The extrapolated relative survival gives a good estimate of the mean survival time in all age groups and all of the four cancer sites (table 1), and performs much better than extrapolation of all-cause survival. For colon cancer and melanoma, cancer sites often seen to reach

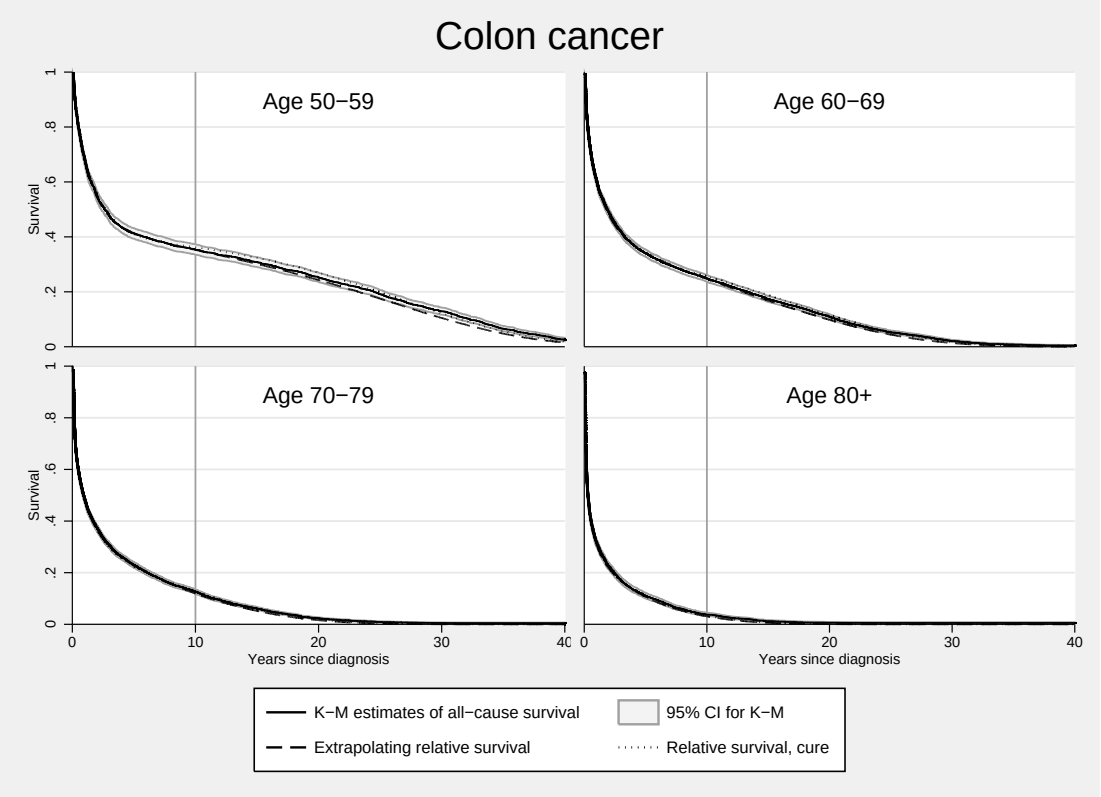

Figure 11 – Observed Kaplan-Meier survival function for colon cancer together with the extrapolated survival from a flexible parametric relative survival model and a flexible parametric cure model.

statistical cure, the extrapolation assuming cure gives good estimates of the mean survival time. Assuming cure does not give good estimates of the mean survival time for bladder cancer or breast cancer, which is not surprising since cure is not a reasonable assumption for these cancer sites. A Weibull distributed excess hazard or a constant excess hazard after the last knot does not yield as good estimates of the mean survival time as the extrapolation of the linear trend from the flexible parametric relative survival model. The mean survival time from the extrapolation of the linear trend are in general very close to the observed mean survival time, most differences are less than 0.5 years. The largest differences are seen for the youngest age group, the group requiring extrapolation over a long period of time. The observed K-M survival function for colon cancer is shown in figure 11 together with the extrapolated survival from a flexible parametric relative survival model as well as a flexible parametric cure model. The three curves agree very well, showing that the extrapolated survival mimics the true observed survival among the cancer patients.

Table 1 – Mean observed survival time (in years) along with the difference between the observed and predicted mean survival times from 5 different extrapolation approaches, by cancer site and age group, diagnosis in Sweden 1961-1970.

Colon cancer	Age group			
	50-59	60-69	70-79	80+
Mean observed survival (years)	10.4	6.32	3.54	1.79
Difference:				
All-cause extrapolated	2.96	2.13	0.71	0.004
Relative survival extrapolated	-0.43	-0.23	-0.20	-0.24
Relative survival, cure	0.10	0.07	-0.13	-0.24
Relative survival, constant excess	-3.03	-1.12	-0.40	-0.26
Relative survival, Weibull distribution	-2.17	-0.95	-0.56	-0.43

Breast cancer	Age group			
	50-59	60-69	70-79	80+
Mean observed survival (years)	14.0	10.0	6.23	3.05
Difference:				
All-cause extrapolated	1.87	1.50	0.34	0.26
Relative survival extrapolated	-0.63	-0.26	-0.05	0.18
Relative survival, cure	1.54	0.88	0.30	0.25
Relative survival, constant excess	-2.66	-0.86	-0.16	0.17
Relative survival, Weibull distribution	-1.97	-0.63	-0.10	-0.15

Melanoma	Age group			
	50-59	60-69	70-79	80+
Mean observed survival (years)	15.9	10.4	5.50	3.22
Difference:				
All-cause extrapolated	3.52	4.60	0.43	0.17
Relative survival extrapolated	-0.95	-0.14	-0.04	-0.05
Relative survival, cure	-0.12	-0.12	0.21	-0.04
Relative survival, constant excess	-3.52	-1.34	-0.24	-0.06
Relative survival, Weibull distribution	-2.61	-1.13	-0.23	-0.04

Bladder cancer	Age group			
	50-59	60-69	70-79	80+
Mean observed survival (years)	13.8	7.94	4.54	2.28
Difference:				
All-cause extrapolated	4.71	1.86	0.39	-0.12
Relative survival extrapolated	0.35	0.05	-0.10	-0.24
Relative survival, cure	1.26	0.58	0.06	-0.20
Relative survival, constant excess	-1.62	-0.63	-0.26	-0.26
Relative survival, Weibull distribution	-0.67	-0.43	-0.30	-0.34

Conclusion

A relative survival approach is useful for extrapolating survival of cancer patients, and the flexible parametric survival model for relative survival can be used to obtain reliable estimates of the mean survival time. This enables estimation of the loss in expectation of life for cancer patients, a measure that has not yet been widely used. The Stata command for flexible parametric survival models `stpm2` [5] has been extended to enable estimation of life expectancy and loss in expectation of life.

6.4 Paper IV

Background

In this study we demonstrated how estimation of loss in expectation of life can be used to address a wide range of research questions concerning the prognosis of cancer patients, using the methods developed in paper III. This is illustrated by investigating life expectancy and loss in expectation of life after a diagnosis of colon cancer, how this changes over calendar time as well as from time since diagnosis, and by quantifying the survival difference between males and females.

Material

A total of 155 851 cases of colon cancer (ICD7 153) were identified in the Swedish Cancer Register during the years 1961–2011, and after restricting to colon adenocarcinoma 135 702 cases were left. Individuals with multiple records of primary colon adenocarcinomas were only included with their first recorded diagnosis which excluded 3 677 diagnoses. We further excluded diagnoses that were detected incidentally at autopsy (4 644 cases), individuals aged less than 20 at diagnosis (38 cases) or if the date of diagnosis was recorded to be after the date of death (39 cases), which left a total of 127 304 patients in the cohort. All patients were followed-up until death or censored at the date of first emigration after diagnosis, 2012-12-31 or 15 years after diagnosis, whichever came first.

Methods

Life expectancy and loss in expectation of life were estimated from a flexible parametric relative survival model as described in section 4.2. Sex, age at diagnosis and year of diagnosis were included in the model along with two-way interactions between them. Age and year were both modelled continuously and non-linearly using restricted cubic splines, with only the linear term included for interactions. Following the results in paper III we performed the extrapolation based on a flexible parametric (relative) survival model without imposing constraints of cure or constant excess hazard, even though a cure model could give a good fit here. To obtain estimates for recently diagnosed patients a separate model including sex and age was fitted using a period analysis approach (see section 2.9) with the period window set to 2007-01-01 – 2012-12-31. Based on the results from the period analysis, the number of life years potentially gained if colon cancer survival among males could be brought to the same level as for females was estimated. This was performed by applying the female cancer mortality rates to males (but keeping the male background mortality rates) and calculating the total amount of life years lost for all patients in 2011 if males had the same colon cancer

survival as females and contrasting this to the total amount of life years lost for all patients in 2011 given the survival differences between males and females. All analyses were performed using the `stpm2` command in Stata (Statacorp, College Station, TX, USA).

Results

The life expectancy for colon cancer patients in Sweden has increased over time (figure 12), but the increase approximately mimics the increase observed in the general population and therefore the loss in expectation of life has only had a modest decrease. As expected, the life expectancy differs greatly by age, with longer life expectancy for younger patients. But since younger patients also have more years to lose, due to a long life expectancy if cancer-free, they on average lose more years due to cancer compared to older colon cancer patients. This gives a complement to the often reported 5-year RSR, since the 5-year RSR is very similar across age for colon cancer patients diagnosed in Sweden during the last 20 years. Female colon cancer patients have a better life expectancy than males, but still have a greater loss in expectation of life, due to higher life expectancy in the general population among females compared to males.

The conditional loss in expectation of life decreased with follow-up time, especially during the first years after diagnosis, as seen in figure 13, and the pattern was similar for males and females and across calendar years. From a patient's perspective, but also for health care planning, conditional estimates are of great importance since the prognosis changes substantially within the first few years after a colon cancer diagnosis.

Based on the period analysis, the total estimated number of life years lost for Swedish colon cancer patients diagnosed in 2011 will be 21 252 years. This number was estimated to be 20 461 if males were given the same colon cancer mortality as females, giving a potential gain of 791 life years. On the individual level this gain would give 0.75 years longer life expectancy for males aged 55 at diagnosis if they had the same cancer patient mortality as females aged 55 at diagnosis, whereas this number is 0.64, 0.36 and 0.11 years for 65, 75 and 85 year old males, respectively.

Sensitivity analysis

A sensitivity analysis was carried out to see how robust the estimates were to the number of knots used for the time-scale as well as for the effect of age and calendar year. This was not included in the paper due to the length requirements of most journals. The main model in this study had 5 degrees of freedom for the baseline log cumulative hazard, 3 degrees of freedom for time-varying effects and both the age and calendar year splines had 4 degrees of freedom. For the sensitivity analysis, four extra models were fitted, one with 4 degrees of

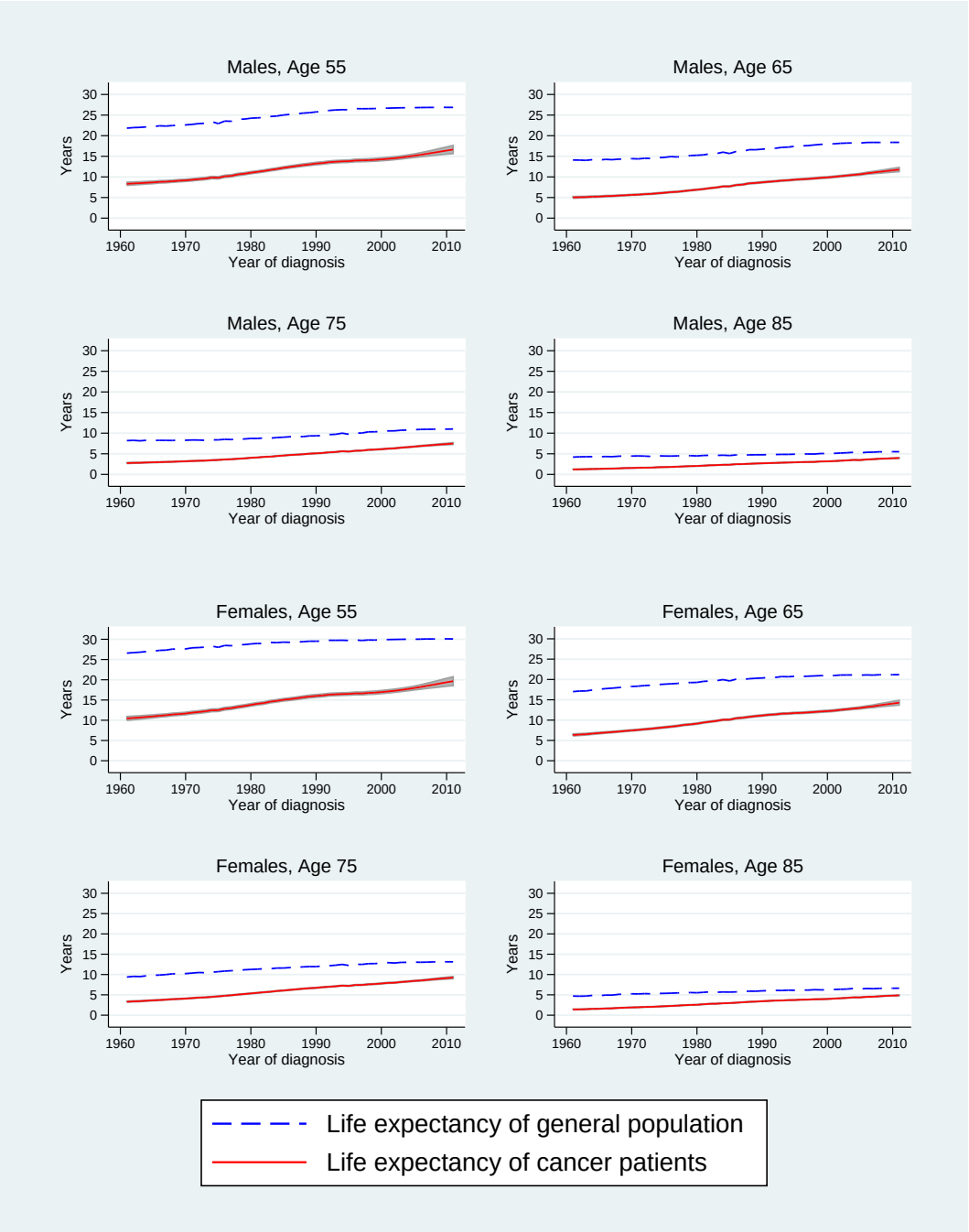

Figure 12 – Temporal trends in life expectancy from diagnosis for colon cancer patients diagnosed in Sweden during 1961–2011.

freedom for the baseline, one with 6 degrees of freedom for the baseline, one with 3 degrees of freedom for the effect of age and calendar year and finally one with 5 degrees of freedom for the effect of age and calendar year. Estimates of the loss in expectation of life from the main analysis and the four sensitivity analysis are presented in figure 14 and 15 for two selected ages and 5 selected years.

The estimates of loss in expectation of life seem to be robust to the number of knots, but

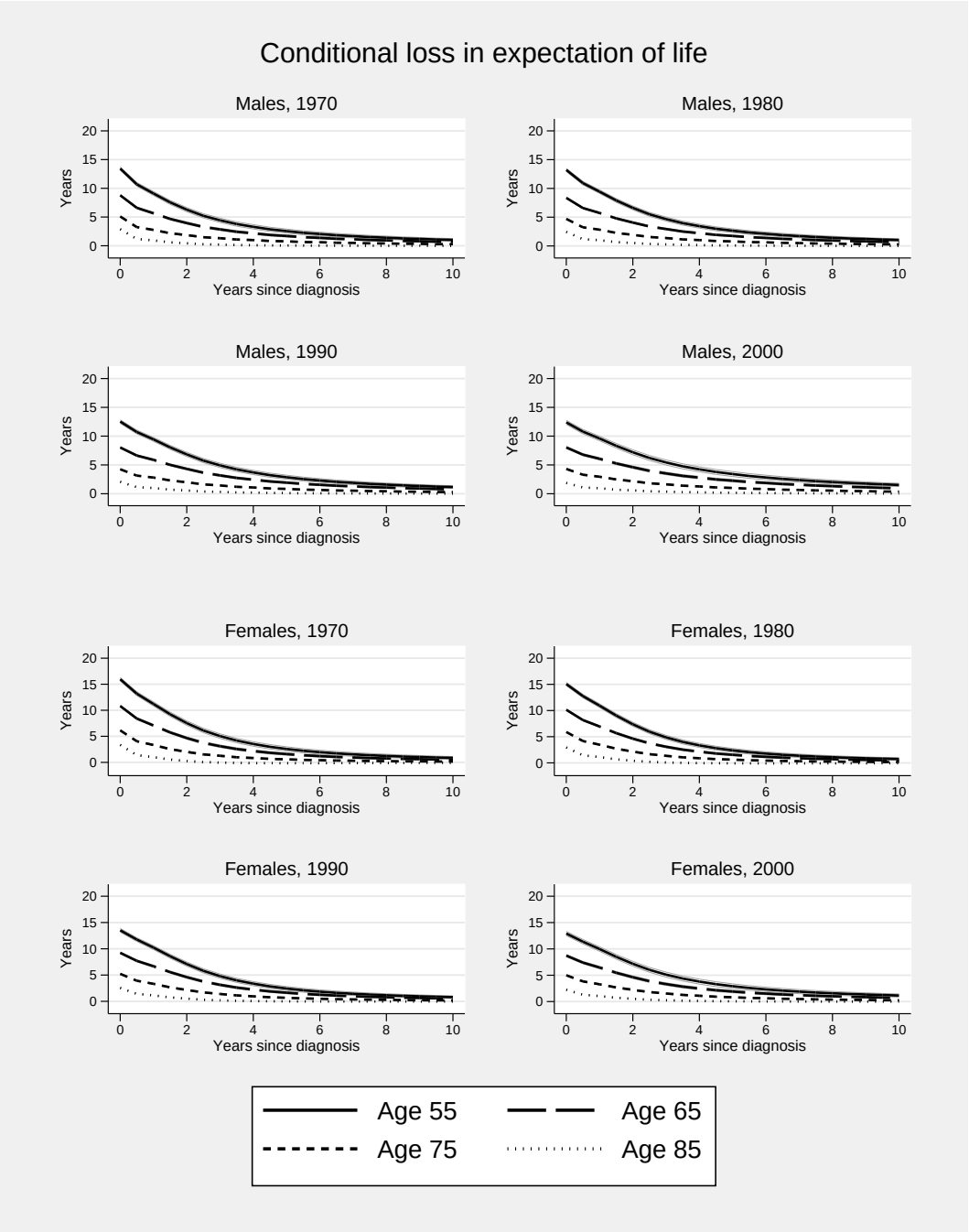

Figure 13 – Loss in expectation of life conditional on time since diagnosis for colon cancer patients diagnosed in Sweden during 1961-2011.

we also wanted to investigate how much small variations impact on the total number of life years lost. Therefore the period analysis was also repeated with 4 degrees of freedom for the baseline, 6 degrees of freedom for the baseline and 5 degrees of freedom for the age effect. The total number of life years lost for patients diagnosed in 2011, estimated from the three sensitivity analyses, ranged from 21 219 to 21 254.

Conclusion

Improved life expectancy was observed for colon cancer patients during the 50 year period under study, but the increase mimicked the improved life expectancy in the general population. Young colon cancer patients have a longer life expectancy, but are also expected to lose more years of their life compared to older colon cancer patients. A similar result was observed for female colon cancer patients who have a longer life expectancy but also lose more years than males. However, for all calendar years, all ages and for both males and females the loss in expectation of life decreases substantially when conditioning on survival up to a few years after diagnosis. In summary, life expectancy and loss in expectation of life improves the understanding of the impact of cancer and gives a good complement to RSR for understanding the prognosis of cancer patients. In this population-based study we demonstrated how summarizing colon cancer survival in terms of loss in expectation of life can be useful in order to gain further insights of the impact of colon cancer on both the individual and population level by examining temporal trends, changes by time since diagnosis and quantifying differences between males and females.

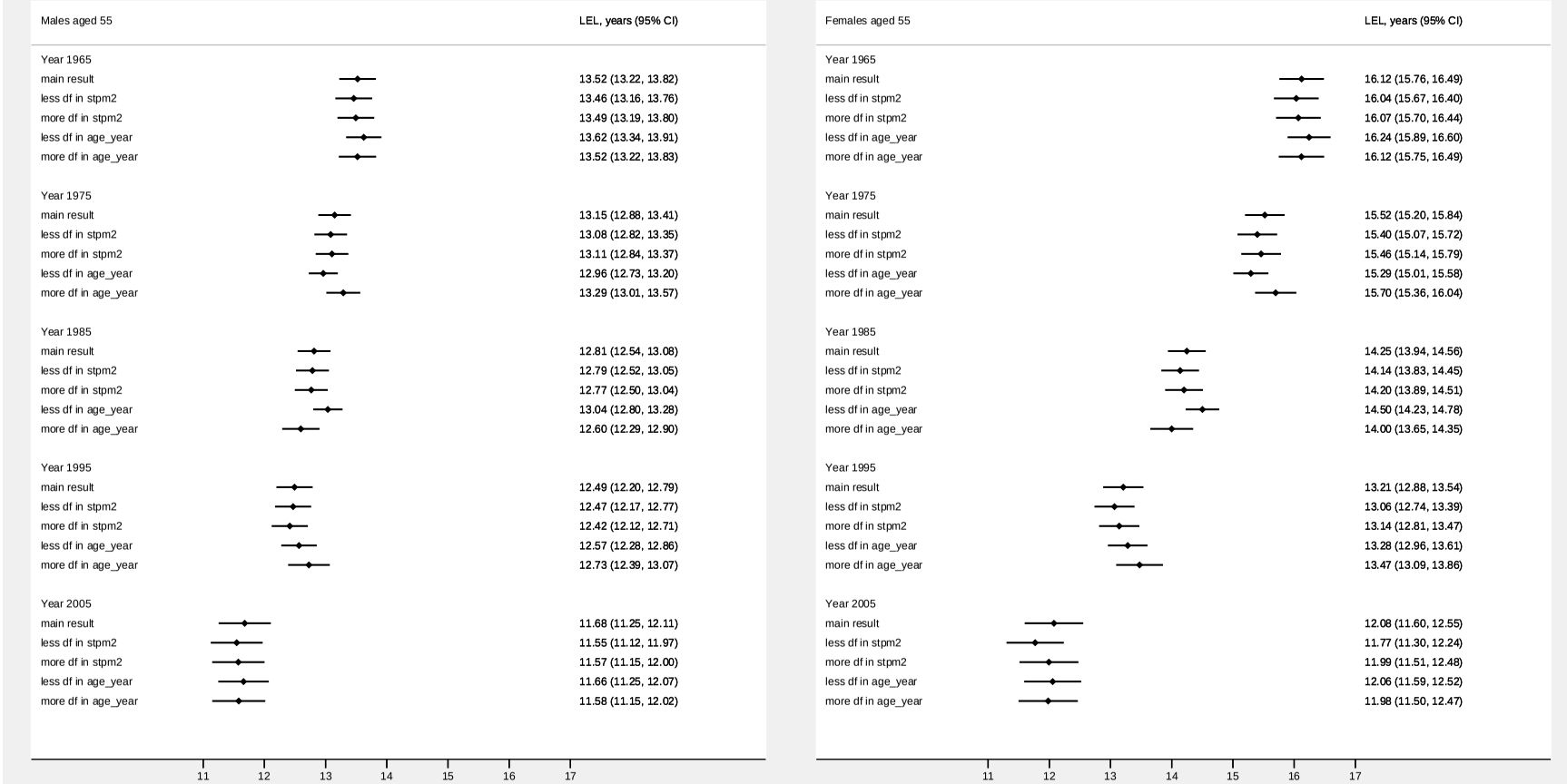

Figure 14 – Sensitivity to knot distribution for estimation of loss in expectation of life.

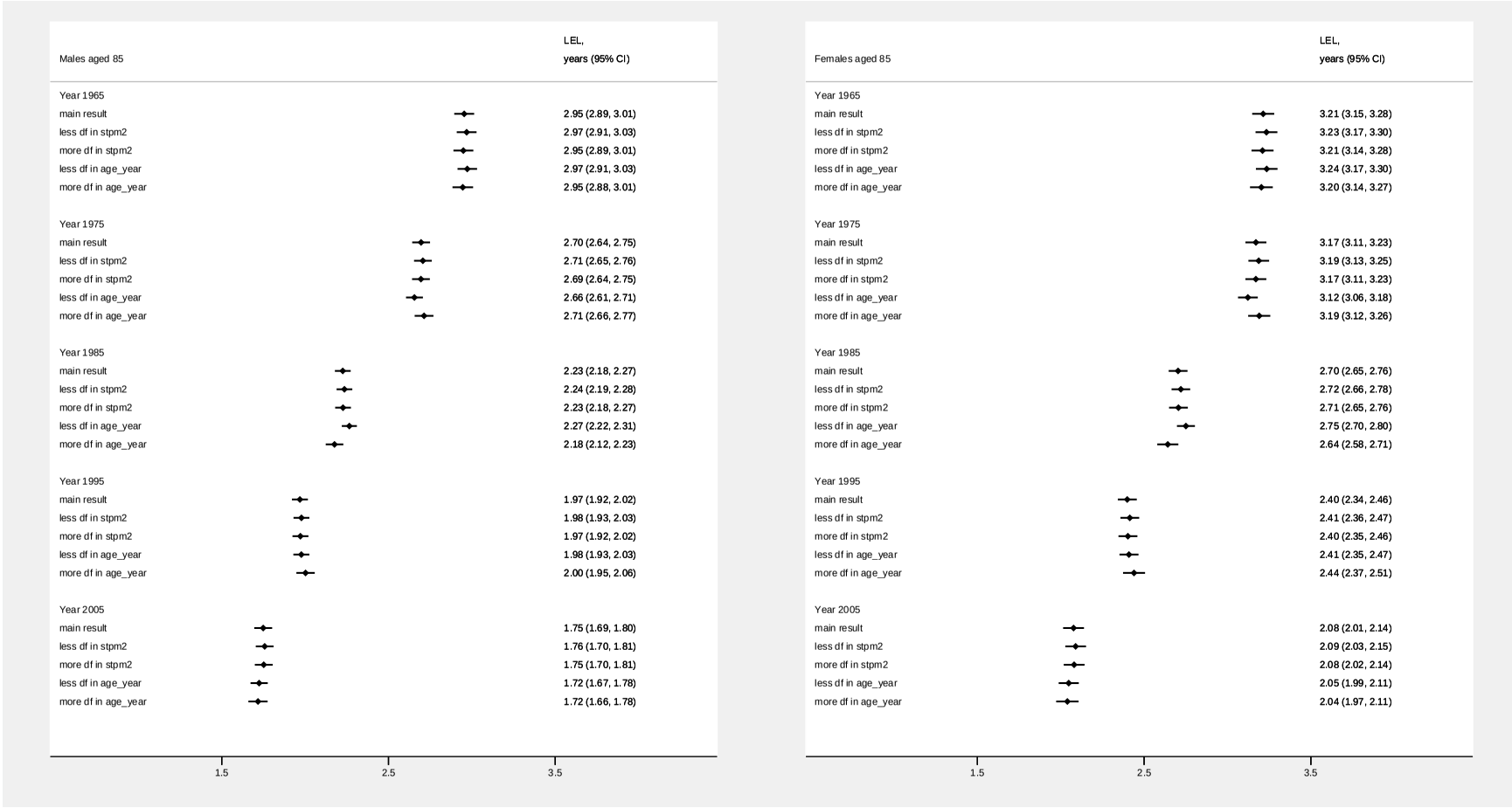

Figure 15 – Sensitivity to knot distribution for estimation of loss in expectation of life.

7 Conclusion and future perspective

Alternative measures of cancer patient survival are needed to gain further insights into cancer patient survival, and in this thesis I have focused on estimation of cure and loss in expectation of life among cancer patients. By the use of different measures, a lot more can be understood about different aspects of the prognosis of cancer patients, and different consumers of survival statistics need different measures to serve their different questions. In paper I the estimation of the proportion cured, along with median survival time of uncured, enhanced the understanding of temporal trends in AML patient survival and the impact of treatment changes over time. In paper IV life expectancy and loss in expectation of life was used as a complement to 5-year RSRs to quantify the survival of colon cancer patients. Paper I and IV were applications of methods for estimating cancer patient survival, and the aim was to gain insights into the prognosis of AML patients and colon cancer patients as well as to present the measures to a broader research community. Paper II and III concerned methods development and extensions, and were published in statistical journals. In paper II a flexible parametric cure model that overcomes some of the limitations of standard parametric cure models was developed. The possibilities of using a flexible parametric relative survival model for extrapolating survival and therefore estimating life expectancy and loss in expectation of life were evaluated in paper III. Evaluations of the methods presented in this thesis were included in the published papers, and selected parts also included in section 6, but more evaluations and further extensions are warranted. A few thoughts on possible extensions and further evaluations are presented in this section.

Flexible parametric cure model

A recent study by Yu *et al.* [85] comparing cure models came to the conclusion that the flexible parametric cure models give valid estimates of cure, and should be preferred over many parametric cure models. What the study did not include was an evaluation of how well the flexible parametric cure model estimates the survival of uncured. This has however been studied in a master thesis from the University of Leicester [86], although the results have not been published. The thesis showed that even though the flexible parametric cure model gives estimates with negligible bias of the cure proportion and the median survival time of uncured there was small bias in the estimation of, for example, the time point where 90% of uncured had died. This is probably due to the restriction put on the splines in the flexible parametric cure model. As described in section 4.3, when the cumulative excess hazard is forced to be constant after the last knot, the shape of the cumulative excess hazard is also slightly changed before the position of the last knot, due to the smoothness of the

splines. This could potentially be improved by the use of the more flexible splines with relaxed smoothness at the last knot, described in the same section. We are planning to conduct a simulation study to investigate if the fit is improved by using the more flexible splines, under different scenarios.

The study by Yu *et al.* also included a sensitivity analysis of the robustness to the choice of knots in the flexible parametric cure model. As described in section 2.6, the flexible parametric survival model has been shown to be robust to the number and location of the knots, but this is not necessarily true when a cure point is imposed. The shape of the baseline log cumulative excess hazard can easily be captured using restricted cubic splines, since the log cumulative excess hazard is a fairly stable function. However, it is not obvious that restricted cubic splines fit the data as well when the spline is manipulated to force a zero slope after the last knot. Another issue is that to give a good estimate of the cure proportion, the splines need to fit the data well towards the end of follow-up, even though most events are observed early in the follow-up. The sensitivity of the flexible parametric cure model to the knot distribution was also examined in paper II. Both sensitivity analyses showed that the flexible parametric cure model is robust to the knot distribution as long as there are enough knots towards the end of follow-up. Therefore, the default knot distribution for the `stpm2` command includes one knot at the 95th as well as the 100th percentile of event times when a flexible parametric cure model is fitted [64].

The flexible parametric cure model is a restricted flexible parametric survival model, where one of the spline parameters is set to 0. Because of this, the flexible parametric cure model is nested within a flexible parametric survival model, and the assumption of cure can therefore be tested using a likelihood ratio (LR) test. This was discussed in paper II, where we also showed the practical limitations of this approach. There is a general problem with the use of LR tests or other likelihood based comparison criteria (such as Akaike information criterion (AIC) and Bayesian information criterion (BIC)) for cure models, since more weight is given to the fit early in follow-up since that is where most events occur. A model with a poorly estimated cure proportion could fit slightly better in the beginning of follow-up and therefore be chosen over a model with a better estimate of cure. An example of this was presented by Lambert *et al.* [39] where a non-mixture model with a log-normal distribution gives a lower AIC than a non-mixture model with Weibull distribution, although it is clear that the Weibull non-mixture model gives a better estimate of the cure proportion. This is especially problematic in the type of population-based studies that are presented in this thesis, since the large data sets with many events will easily give significant results even for very small differences. When using a cure model we are generally willing to give up some model fit early on in follow-up in order to gain the extra information that a cure model

can provide. We would therefore often chose to use a cure model even if a LR test would suggest that a model without a cure point gives a slightly better fit. We recommend that the existence of a cure point is evaluated by plotting empirical life table estimates of relative survival, and that cure models should not be used if the relative survival does not reach a plateau within the available follow-up. A log-rank type test for testing the existence of cure in a relative survival setting has also been proposed [87]. A possible further extension of the flexible parametric cure model could be to use a likelihood approach that gives more weight to the end of follow-up, and by that enable comparison to a flexible parametric survival model without cure.

A difference between the flexible parametric cure model and the other cure models described in this thesis is that the point of cure is specified in the flexible parametric cure model by the position of the final knot. When the cure proportion is estimated as an asymptote of the survival function, as for the mixture and non-mixture cure model, the cure point is effectively at infinity. This feature of the cure models could be seen as both an advantage and a disadvantage. By explicitly specifying the point of cure in the flexible parametric cure model (by the position of the last knot), the cure proportion could be grossly overestimated if the last knot is placed too early, if the survival function has not yet reached a plateau. On the other hand, estimating the cure point itself could be of interest and by comparing models with the last knot placed at different time points one could potentially get an estimate of the time of cure. This however needs further work and evaluation. Even though it could theoretically be done, in practice it is difficult to compare models in this way. Comparisons of cure models with the last knot at different positions is difficult for the reasons described above, another reason is that the flexible parametric survival model is robust to the positions of the knots, so small changes in the knot position will only lead to small changes in model fit and therefore this approach is unlikely to be useful for predicting time of cure.

I believe that cure models are a very useful tool, and hope that the flexible parametric cure model will lead to more use of the cure proportion as an estimate of long-term survival. The main advantages of the flexible parametric cure model is that it allows inclusion of patient groups for which parametric cure models do not converge or do not give a good fit. We have, for example, used the flexible parametric cure model to study how the cure proportion and median survival time of uncured differs by age, stage, location and sex for patients with malignant melanoma. This was not possible to do with a Weibull mixture or non-mixture cure model because the models wouldn't converge. Malignant melanoma is a cancer with relatively high survival, and in our experience the Weibull mixture and non-mixture cure models do not always converge when the relative survival is high. In summary, the flexible parametric cure model is a good alternative to the mixture and non-mixture cure

models, but further evaluations are needed.

Life expectancy and loss in expectation of life

The loss in expectation of life is theoretically easy to estimate, but in practice there are complications because of limited follow-up. Therefore, the key feature of a method used to estimate the loss in expectation of life is that the observed survival can be satisfactorily extrapolated beyond the available follow-up. In paper III it was shown that a flexible parametric relative survival model can be used, and the extrapolated survival agreed very well with the observed survival. In that study all patients diagnosed at ages younger than 50 were excluded, since the available 40 years of follow-up was not enough to estimate their mean survival time. The extrapolation might not perform as well for younger patients both since it is a smaller group and since the extrapolation needs to be done over a longer period, and it has also been argued that long-term excess mortality among patients diagnosed at a young age may have a large impact on mortality in later ages [88]. Further evaluations of extrapolations of the flexible parametric relative survival model are therefore needed in order to know how well they can be applied for estimating the life expectancy of younger cancer patients. It would also be interesting to see how well the models extrapolate survival in smaller data sets such as clinical trials, and if this approach could be a complement to methods often used to compare treatment groups. Because of the limited follow-up, restricted mean survival times are often presented, and Royston and Parmar [89] suggested the use of flexible parametric models. Andersen suggested a way of dividing the number of life years lost within a restricted time interval according to different causes of death using competing risks methods [90]. However, it would be interesting to calculate the full life expectancy rather than the restricted, since differences in the restricted mean survival between treatment groups can never fully estimate the impact of treatment differences on survival. Another potential area where this approach could be used is for economic evaluations and cost-effectiveness studies. Tappenden *et. al* argue that alternative extrapolation methods are needed in cost-effectiveness studies [91]. Some methods have been proposed, such as poly-Weibull models [92] or Bayesian evidence synthesis of data from a variety of sources [93], but evaluations of these methods have not been performed on data with complete follow-up.

Not only the observed survival, but also the expected survival has to be extrapolated. In both paper III and IV we assumed that the population mortality rates stayed constant after the last available calendar year in the population mortality files (1980 in paper III and 2011 in paper IV). Even so, the extrapolations made in paper III agreed very well with the observed mean survival time. Statistics Sweden has published projections of the Swedish population mortality [94] , and when available one should consider using projections

Table 2 – Estimated life years lost for colon cancer patients diagnosed in Sweden during 2011, using two different assumptions about future population mortality rates.

	Estimated life years lost from paper IV assuming constant population mortality rates after 2011		Estimated life years lost using projection of future population mortality rates from Statistics Sweden	
Age	Males	Females	Males	Females
55	10.7	11.3	12.2	12.3
65	7.12	7.66	8.02	8.40
75	4.00	4.50	4.40	4.88
85	1.77	2.05	1.86	2.18
Total in population	10580	10672	11843	11637

of population mortality when future mortality is needed. We reproduced the predictions from the period analysis in paper IV using the future population mortality projections from Statistics Sweden, and the results are shown in table 2. The estimated loss in expectation of life is higher if projections of future population mortality rates are used, since the mortality rate in Sweden is expected to be decreasing. Even so, one has to keep in mind that the projections are also estimates.

Another issue with the use of population mortality rates is if the expected rates can be considered to be fixed, or if they should be considered random. In many applications of relative survival the expected survival can be considered fixed since it is based on such a large population so that random variation is negligible. However, if the population mortality file is broken down by many factors or is specific for a small region one has to start considering if the assumption is still valid. Seppä *et al.* incorporated the uncertainty of the population mortality into their estimation in a study focusing on regional variation in cancer patient survival in Finland [95], but it did not have a large impact on the estimates or their standard errors. Another reason for taking the variation of the population mortality into account is when estimates of the expected survival are of specific interest and not only for estimation of relative survival. This is the case when estimating the loss in expectation of life, since it requires the estimation of life expectancy in the absence of cancer, obtained from population mortality rates. For the estimation of the loss in expectation of life, we only consider the life expectancy in the presence of cancer as random and assume that we can correctly estimate the life expectancy in the absence of cancer without any variation. In order to understand how large the impact of this is on the variance of the loss in expectation of life, further analysis is needed where the variation in the life expectancy in the absence of cancer is taken into account. Another, and easier, approach to reduce the variation in the population mortality would be to smooth the population mortality rates. Although, this approach still

assumes that there is no random variation in the estimation of expected survival for the cancer patients.

A potential important use of the loss in expectation of life is to quantify differences between groups. It is of major public health interest to be able to quantify differences between, for example, socio-economic groups, and the number of life years lost is one possible measure which is easily interpreted. During the last few years there has been a great interest in quantifying differences in cancer patient survival between countries in other ways than by comparing the 5- or 10-year RSR, and the number of avoidable deaths has been used [96, 97, 98, 99]. A disadvantage of avoidable deaths is that the measure is highly time-dependent, since deaths can only be postponed and not avoided in the long run. We believe that this is an area where the loss in expectation of life could be used instead to better quantify the impact of differences in cancer patient survival between countries or between groups in the population.

Final words

In summary, the methods presented in this thesis are additional tools for estimating and quantifying population-based cancer patient survival using routinely collected cancer registry data. They can hopefully lead to a further understanding of different aspects of cancer patient survival, and how cancer patient survival differs between groups. By the implementation of the methods in user-friendly software we have enabled others to use the methods, and applications of the methods will hopefully increase the awareness of these and other alternative measures of cancer patient survival.

8 Acknowledgements

First I would like to thank my wonderful supervisors! My main supervisor **Paul Lambert** for always taking time for me when I need it, but also for letting me do things on my own and in my own time when that's what I've wanted. Thanks to my co-supervisor and former boss **Paul Dickman** for opening the door to research for me, and always inspiring me with your great passion for science. And **Mats Lambe**, my second co-supervisor, thank you for your support, your great input to my manuscripts and for always being calm and calming me down when I'm stressed. I've enjoyed working with the three of you and hope that we will continue collaborating in the future!

Sandra Eloranta, my partner in crime, what would I have done without you? Thanks for our endless talks, keeping me company on all our trips, the morning walks in Leicester, for all the input to my work and for teaming up with me against our supervisors during the lively discussions.

A big thanks to my friends and colleagues **Anna Johansson** and **Caroline Weibull**, for always making me want to go to work. MEB would not be the same without you.

I would like to thank the head of department **Henrik Grönberg** and the previous heads of the department **Nancy Pedersen** and **Hans-Olov Adami** for making MEB such a wonderful place with a great atmosphere, and a perfect environment for a young researcher to be brought up in.

A special thanks to all my co-authors, **Åsa Rangert Derolf**, **Sigurdur Yngvi Kristinsson**, **Ola Landgren**, **Magnus Björkholm** and **Annika Sjövall**. Thank you all for great collaboration! That also goes for the rest of the **HEMEP** group, I hope we will continue our collaboration and cake eating. I would also like to thank **Johan Hansson**, **Hanna Eriksson** and **Eva Månsson-Brahme**, co-authors on a paper that did not end up in this thesis.

I would like to thank present and former members in the biostatistics group at MEB for making this a fantastic work place. Also thanks to my present and former fellow PhD-students and all other friends and colleagues at MEB. Keep up the MEB spirit! A special thanks to **Marie Jansson** and **Camilla Ahlqvist** for all the help with administration. Thanks to **Robert Karlsson** for giving feedback on my kappa.

To all my friends in Leicester, especially **Sally Hinchliffe**, **Mark Rutherford** and **Michael Crowther**, thank you for taking care of me during my visits!

To all former participants on the course "Statistical methods for population-based cancer survival analysis", thank you for inspiring discussions and for teaching me so much.

Last but not least, a big thank you to my husband, my family and my friends!

9 References

[1] Dickman PW, Adami HO. Interpreting trends in cancer patient survival. *J Intern Med* 2006;**260**:103–117.

[2] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Statistics in Medicine* 2004;**23**:51–64.

[3] Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002;**21**:2175–2197.

[4] Royston P. Flexible parametric alternatives to the cox model, and more. *The Stata Journal* 2001;**1**:1–28.

[5] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal* 2009;**9**:265–290.

[6] Smith PL. Splines as a useful and convenient statistical tool. *The American Statistician* 1979;**33**:57–62.

[7] Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989;**8**:551–561.

[8] Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 2007; **26**:5486–5498.

[9] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin* 2011;**61**:69–90.

[10] The National Board of Health and Welfare. Causes of death 2012. Tech. rep., The National Board of Health and Welfare, 2013.

[11] The National Board of Health and Welfare. Cancer incidence in Sweden 2011. Tech. rep., The National Board of Health and Welfare, 2012.

[12] Muir CS. The cancer registry in cancer control: An overview. IARC Scientific Publications No. 66. Lyon: International Agency for Research on Cancer, 1985; 13–26.

[13] Armstrong BK. The role of the cancer registry in cancer control. *Cancer Causes and Control* 1992;**3**:569–579.

[14] Parkin DM. The role of cancer registries in cancer control. *Int J Clin Oncol* 2008; **13**:102–111.

[15] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007;**26**:2389–2430.

[16] Klein J, Moeschberger ML. *Survival analysis: tecniques for censored and truncated data 2nd edition*. Springer, 2003.

[17] Ederer F, Axtell L, Cutler S. The relative survival rate: A statistical methodology. *National Cancer Institute Monograph* 1961;**6**:101–121.

[18] The human mortality database. http://www.mortality.org/.

[19] Wilmoth J, Andreev K, Jdanov D, Glei D. Methods protocol for the human mortality database. Technical report Version 5, 2007.

[20] Fall K, Strömberg F, Rosell J, Andrèn O, Varenhorst E, South-East Region Prostate Cancer Group. Reliability of death certificates in prostate cancer patients. *Scand J Urol Nephrol* 2008;**42**:352–357.

[21] Begg CB, Schrag D. Attribution of deaths following cancer treatment. *J Natl Cancer Inst* 2002;**94**:1044–1045.

[22] Pohar Perme M, Stare J, Estève J. On estimation in relative survival. *Biometrics* 2012; **68**:113–120.

[23] Roche L, Danieli C, Belot A, Grosclaude P, Bouvier AM, Velten M, Iwaz J, Remontet L, Bossard N. Cancer net survival on registry data: Use of the new unbiased pohar-perme estimator and magnitude of the bias with the classical methods. *Int J Cancer* 2012; **132**:2359–69.

[24] Rutherford MJ, Dickman PW, Lambert PC. Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiology* 2012;**36**:16–21.

[25] Dickman PW, Lambert PC, Coviello E, Rutherford MJ. Estimating net survival in population-based cancer studies. *Int J Cancer* 2013;**133**:519–21.

[26] Hakulinen T, Tenkanen L. Regression analyses of relative survival rates. *Applied Statistics* 1987;**36**:309–317.

[27] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med* 1990;**9**:529–538.

[28] Bolard P, Quantin C, Abrahamowicz M, Esteve J, Giorgi R, Chadha-Boreham H, Binquet C, Faivre J. Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *J Cancer Epidemiol Prev* 2002;**7**:113–122.

[29] Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, Faivre J. A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine* 2003;**22**:2767–2784.

[30] Sasieni PD. Proportional excess hazards. *Biometrika* 1996;**83**:127–141.

[31] Pohar Perme M, Henderson R, Stare J. An approach to estimation in relative survival regression. *Biostatistics* 2009;**10**:136–146.

[32] Stata. *Stata base reference manual Release 13*. Stata Press: College Station, 2013.

[33] Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation* 2013;(in press).

[34] Eloranta S, Lambert PC, Sjöberg J, Andersson TML, Björkholm M, Dickman PW. Temporal trends in mortality from diseases of the circulatory system after treatment for hodgkin lymphoma: a population-based cohort study in sweden (1973 to 2006). *J Clin Oncol* 2013;**31**:1435–1441.

[35] Boag J. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society B* 1949;**11**:15–44.

[36] Berkson J, Gage R. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952;**47**:501–515.

[37] Verdecchia A, Angelis RD, Capocaccia R, Sant M, Micheli A, Gatta G, Berrino F. The cure for colon cancer: results from the EUROCARE study. *International Journal of Cancer* 1998;**77**:322–329.

[38] De Angelis R, Capocaccia R, Hakulinen T, Söderman B, Verdecchia A. Mixture models for cancer survival analysis: Application to population-based data with covariates. *Statistics in Medicine* 1999;**18**:441–454.

[39] Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007;**8**:576–594.

[40] Yu B, Tiwari RC, Cronin KA, Feuer EJ. Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine* 2004;**23**:1733–1747.

[41] Yu B, Tiwari RC, Cronin KA, McDonald C, Feuer EJ. Cansurv: A windows program for population-based cancer survival analysis. *Comput Methods Programs Biomed* 2005; **80**:195–203.

[42] Lambert PC, Dickman PW, Österlund P, Andersson T, Sankila R, Glimelius B. Temporal trends in the proportion cured for cancer of the colon and rectum: a population-based study using data from the Finnish cancer registry. *International Journal of Cancer* 2007; **121**:2052–2059.

[43] Francisci S, Capocaccia R, Grande E, Santaquilani M, Simonetti A, Allemani C, Gatta G, Sant M, Zigon G, Bray F, Janssen-Heijnen M, EUROCAREWG. The cure of cancer: a european perspective. *Eur J Cancer* 2009;**45**:1067–1079.

[44] Eloranta S, Lambert PC, Cavalli-Björkman N, Andersson TML, Glimelius B, Dickman PW. Does socioeconomic status influence the prospect of cure from colon cancer–a population-based study in Sweden 1965-2000. *Eur J Cancer* 2010;**46**:2965–2972.

[45] Sposto R. Cure model analysis in cancer: an application to data from the children's cancer group. *Statistics in Medicine* 2002;**21**:293–312.

[46] Yakovlev AY, Tsodikov A. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, 1996.

[47] Tsodikov AD, Ibrahim JG, Yakovlev AY. Estimating cure rates from survival data: An alternative to two-component mixture models. *J Am Stat Assoc* 2003;**98**:1063–1078.

[48] Andrae B, Andersson TML, Lambert PC, Kemetli L, Silfverdal L, Strander B, Ryd W, Dillner J, Törnberg S, Sparén P. Screening and cervical cancer cure: population based cohort study. *BMJ* 2012;**344**:e900.

[49] Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer* 1996;**78**:2004–2010.

[50] Brenner H, Gefeller O. Deriving more up-to-date estimates of long-term patient survival. *Journal of Clinical Epidemiology* 1997;**50**:211–216.

[51] Talbäck M, Stenbeck M, Rosén M. Up-to-date long-term survival of cancer patients: an evaluation of period analysis on Swedish Cancer Registry data. *Eur J Cancer* 2004; **40**:1361–1372.

[52] Talbäck M, Dickman PW. Predicting the survival of cancer patients recently diagnosed in sweden and an evaluation of predictions published in 2004. *Acta Oncol* 2012;**51**:17–27.

[53] Lambert PC. Modeling of the cure fraction in survival studies. *The Stata Journal* 2007; **7**:351–375.

[54] Lambert PC, Dickman PW, Weston CL, Thompson JR. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society Series C* 2010;**59**:35–55.

[55] Taylor JMG. Semi-parametric estimation in failure time mixture models. *Biometrics* 1995;**51**:899–907.

[56] Tsodikov A. A proportional hazards model taking account of long-term survivors. *Biometrics* 1998;**54**:1508–1516.

[57] Peng Y, Dear KBG. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000;**56**:237–243.

[58] Sy JP, Taylor JMG. Estimation in a cox proportional hazards cure model. *Biometrics* 2000;**56**:227–236.

[59] Kim S, Chen MH, Dey DK. A new threshold regression model for survival data with a cure fraction. *Lifetime Data Anal* 2011;**17**:101–122.

[60] Rodrigues J, Cancho VG, de Castro M, Balakrishnan N. A bayesian destructive weighted poisson cure rate model and an application to a cutaneous melanoma data. *Stat Methods Med Res* 2012;**21**:585–597.

[61] Corbiere F, Commenges D, Taylor JMG, Joly P. A penalized likelihood approach for mixture cure models. *Stat Med* 2009;**28**:510–524.

[62] Myasnikova E. Spline-based estimation of cure rates: an application to the analysis of breast cancer data. *Mathematical and computer modelling* 2000;**32**:217–228.

[63] Wang L, Du P, Liang H. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics* 2012;**68**:726–735.

[64] Andersson TML, Lambert PC. Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models. *The Stata Journal* 2012;**12**:623–628.

[65] Viscomi S, Pastore G, Dama E, Zuccolo L, Pearce N, Merletti F, Magnani C. Life expectancy as an indicator of outcome in follow-up of population-based cancer registries: the example of childhood leukemia. *Annals of Oncology* 2006;**17(1)**:167–171.

[66] Brown ML, Lipscomb J, Snyder C. The burden of illness of cancer: economic cost and quality of life. *Annu Rev Public Health* 2001;**22**:91–113.

[67] Hakama M, Hakulinen T. Estimating the expectation of life in cancer survival studies with incomplete follow-up information. *Journal of Chronic Diseases* 1977;**30**:585–597.

[68] Messori A, Trippoli S. A new method for expressing survival and life expectancy in lifetime cost-effectiveness studies that evaluate cancer patients (review). *Oncol Rep* 1999;**6**:1135–1141.

[69] Hwang JS, Wang JD. Monte carlo estimation of extrapolation of quality-adjusted survival for follow-up studies. *Stat Med* 1999;**18**:1627–1640.

[70] Chu PC, Wang JD, Hwang JS, Chang YY. Estimation of life expectancy and the expected years of life lost in patients with major cancers: extrapolation of survival curves under high-censored rates. *Value Health* 2008;**11**:1102–1109.

[71] Nelson CL, Sun JL, Tsiatis AA, Mark DB. Empirical estimation of life expectancy from large clinical trials: use of left-truncated, right-censored survival analysis methodology. *Stat Med* 2008;**27**:5525–5555.

[72] Beck JR, Kassirer JP, Pauker SG. A convenient approximation of life expectancy (the "DEALE"). I. validation of the method. *Am J Med* 1982;**73**:883–888.

[73] Gelber RD, Goldhirsch A, Cole BF. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. international breast cancer study group. *Control Clin Trials* 1993;**14**:485–499.

[74] Caro JJ, Ishak KJ, Migliaccio-Walle K. Estimating survival for cost-effectiveness analyses: a case study in atherothrombosis. *Value Health* 2004;**7**:627–635.

[75] Straatman H, Verbeek ALM, Peer PGM, Borm G. Estimating life expectancy and related probabilities in screen-detected breast cancer patients with restricted follow-up information. *Stat Med* 2004;**23**:431–448.

[76] Horm JW, Sondik EJ. Person-years of life lost due to cancer in the united states, 1970 and 1984. *Am J Public Health* 1989;**79**:1490–1493.

[77] Mettlin C. Trends in years of life lost to cancer: 1970-1985. *CA Cancer J Clin* 1989; **39(1)**:33–39.

[78] Burnet NG, Jefferies SJ, Benson RJ, Hunt DP, Treasure FP. Years of life lost (YLL) from cancer is an important measure of population burden – and should be considered when allocating research funds. *Br J Cancer* 2005;**92**:241–245.

[79] Stoer J, Bulirsch R. *Introduction to Numerical Analysis (3rd ed.)*. Springer, 2002.

[80] Pawitan Y. *In all likelihood – Statistical modelling and inference using likelihood*. Oxford University Press, 2001.

[81] Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish cancer register: a sample survey for year 1998. *Acta Oncol* 2009;**48**:27–33.

[82] Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. experience in finland. *Acta Oncol* 1994;**33**:365–369.

[83] Derolf ÅR, Kristinsson SY, Andersson TML, Landgren O, Dickman PW, Björkholm M. Improved patient survival for acute myeloid leukemia: a population-based study of 9729 patients diagnosed in Sweden between 1973 and 2005. *Blood* 2009;**113**:3666–3672.

[84] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958;**53**:457–481.

[85] Yu XQ, De Angelis R, Andersson TML, Lambert PC, O'Connell DL, Dickman PW. Estimating the proportion cured of cancer: Some practical advice for users. *Cancer Epidemiol* 2013;.

[86] Seaton S. Cure models. *MSc in Medical Statistics Thesis* MSc in Medical Statistics Thesis, 2011;**Department of Health Sciences**:University of Leicester.

[87] Yu B. A minimum version of log-rank test for testing the existence of cancer cure using relative survival data. *Biom J* 2012;**54**:45–60.

[88] Reulen RC, Winter DL, Frobisher C, Lancashire ER, Stiller CA, Jenney ME, Skinner R, Stevens MC, Hawkins MM, BCCSSSG. Long-term cause-specific mortality among survivors of childhood cancer. *JAMA* 2010;**304**:172–179.

[89] Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med* 2011;**30**:2409–2421.

[90] Andersen PK. Decomposition of number of life years lost according to causes of death. *Stat Med* 2013;.

[91] Tappenden P, Chilcott J, Ward S, Eggington S, Hind D, Hummel S. Methodological issues in the economic analysis of cancer treatments. *Eur J Cancer* 2006;**42**:2867–2875.

[92] Demiris N, Lunn D, Sharples LD. Survival extrapolation using the poly-weibull model. *Stat Methods Med Res* 2011;.

[93] Demiris N, Sharples LD. Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies. *Stat Med* 2006;**25**:1960–1975.

[94] Statistics Sweden. The future population of sweden 2012-2060. Demographic reports 2012:2, Statistics Sweden, 2012.

[95] Seppä K, Hakulinen T, Kim HJ, Läärä E. Cure fraction model with random effects for regional variation in cancer survival. *Stat Med* 2010;**29**:2781–2793.

[96] Abdel-Rahman M, Stockton D, Rachet B, Hakulinen T, Coleman MP. What if cancer survival in Britain were the same as in Europe: how many deaths are avoidable? *British Journal of Cancer* 2009;**101 Suppl 2**:S115–S124.

[97] Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, Nur U, Tracey E, Coory M, Hatcher J, McGahan CE, Turner D, *et al.*. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the international cancer benchmarking partnership): an analysis of population-based cancer registry data. *Lancet* 2011;**377**:127–138.

[98] Morris EJA, Sandin F, Lambert PC, Bray F, Klint Å, Linklater K, Robinson D, Påhlman L, Holmberg L, Møller H. A population-based comparison of the survival of patients with colorectal cancer in England, Norway and Sweden between 1996 and 2004. *Gut* 2011;**60**:1087–1093.

[99] Lambert PC, Holmberg L, Sandin F, Bray F, Linklater KM, Purushotham A, Robinson D, Møller H. Quantifying differences in breast cancer survival between England and Norway. *Cancer Epidemiology* 2011;**35**:526–533.