

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

Development and Application of Statistical Methods for
Population-Based Cancer Patient Survival

Sandra Eloranta



**Karolinska
Institutet**

Stockholm 2013

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

Published by Karolinska Institutet. Printed by Universitetservice AB.

© Sandra Eloranta, 2013

ISBN 978-91-7549-298-8

Methodological statistical analysis is a necessary beginning of any study of this sort if the study is to be scientifically enlightening.

But it is not the end;

it will have to be subject to judgements based on knowledge of the material that goes beyond the purely statistical analysis in this study as in all others.

JOSEPH BERKSON, 1942

Abstract

The overarching aim of this work has been to develop and apply statistical methods for estimating cancer patient survival from population-based register data. Particular focus has been on statistical methods that can be used for presenting cancer survival statistics from administrative health data registers in a manner that is relevant for physicians and patients.

Study 1: In this study we clarify and discuss the relative merits of estimates of crude and net cancer patient survival, respectively. In addition, we demonstrate how period analysis, applied in a competing risks setting, can be utilised to predict crude survival probabilities applicable to newly diagnosed cancer patients. As a motivating clinical example, we use data from the National Prostate Cancer Register to assess the impact of prognostic factors on the risk of prostate cancer death in relation to death from other causes than prostate cancer, and event-free survival, among recently diagnosed patients. We conclude that the period estimates of crude survival offer a useful basis for risk communication between physicians and clinicians and advocate their use as means to answer prognostic questions.

Study 2: Late adverse health effects in cancer patients are a growing problem given the longer survival seen for most cancers. Deaths that occur as a consequence of treatment toxicity can be regarded as indirect deaths due to cancer. In this methodological study we extend flexible parametric survival models for relative survival by partitioning the overall excess mortality from cancer into two component parts; excess mortality from diseases of the circulatory system, DCS, (assumed caused by the treatment), and remaining excess cancer mortality. We present summary measures for quantifying the risk for death from late effects of treatment relative to the overall risk of dying of breast cancer, or causes unrelated to the cancer. The method is illustrated using data obtained from the Swedish Cancer Register on women diagnosed with breast cancer in Sweden between 1973 and 1992.

Study 3: Survival after Hodgkin lymphoma has increased substantially in the past four decades, following the development of effective multi-agent chemotherapy, introduction of combined-modality therapy with reductions in radiation field size and dose, and more apt evaluation of treatment response. The aim of this study was to present clinically interpretable estimates of temporal trends in the burden of fatal excess DCS mortality among Hodgkin lymphoma survivors who were treated in the 1970's through 1990's, and to predict the future clinical burden among patients diagnosed more recently. Using data from the Swedish Cancer Registry we showed how the excess DCS mortality, within 20 years after diagnosis, has decreased continually since the mid-1980s and is expected to further decrease among patients diagnosed in the modern era. However, when accounting for competing causes of death, we found that excess DCS mortality constitutes a relatively small proportion of the overall mortality among Hodgkin lymphoma patients in Sweden.

Study 4: In this study we show how recently developed flexible parametric cure models, combined with competing risks theory, can be used to estimate crude probabilities that cancer patients who are alive will eventually die from their cancer, or from other causes, respectively. Moreover, we show how to 'update' the prognosis for patients who have survived some time after their diagnosis via the use of conditional probabilities. The method is discussed and demonstrated using data from the Swedish Cancer Register on patients diagnosed with melanoma, colon cancer and acute myeloid leukemia between 1973 and 2007.

List of publications

- I. Eloranta S, Adolfsson J, Lambert PC, Stattin P, Akre O, Andersson TM-L, Dickman PW
How can we make survival statistics more useful for patient and clinicians - an illustration using localized prostate cancer in Sweden.
Cancer Causes & Control 2013; 24(3):505-15.
- II. Eloranta S, Lambert PC, Andersson TM-L, Czene K, Hall P, Björkholm M, Dickman PW
Partitioning of excess mortality in population-based cancer patient survival studies using flexible parametric survival models.
BMC Medical Research Methodology 2012; 24;12(1):86.
- III. Eloranta S, Lambert PC, Sjöberg J, Andersson TM-L, Björkholm M, and Dickman PW
Temporal trends in mortality from diseases of the circulatory system after treatment for Hodgkin lymphoma - A population-based cohort study in Sweden (1973-2006).
Journal of Clinical Oncology 2013; 10;31(11):1435-41.
- IV. Eloranta S, Lambert PC, Andersson TM-L, Björkholm M, and Dickman PW
The application of cure models in the presence of competing risks - a tool for improved risk communication in population-based cancer patient survival studies.
Manuscript.

The studies in this thesis were supported by grants from the Swedish Cancer Society (Studies I-IV) and the Adolf H Lundin Charitable Foundation (Studies III-IV).

Contents

1	Introduction	1
2	Aims of the thesis	3
3	Background	4
3.1	Cancer	4
3.2	The role of a cancer registry in cancer control	5
3.3	Outcome measures used in cancer control	6
3.4	Communicating and producing statistics relevant for clinicians and patients	6
4	Material	9
4.1	Cancer registration in Sweden	9
4.2	Data material used in this thesis	9
4.2.1	Cancer data	9
4.2.2	Death data	10
4.2.3	Population Statistics	11
4.3	A note on record linkage	11
5	Statistical methods	12
5.1	Relative survival and excess mortality	12
5.1.1	Modelling excess mortality using flexible parametric survival models	13
5.2	Period analysis	15
5.3	Statistical methods for competing risks	17
5.3.1	Numerical integration and variance approximation	18
5.4	Cure	20
5.4.1	Modelling statistical cure	21
5.4.2	Flexible parametric cure models	23
6	Study 1: Making cancer patient survival statistics more useful for patients and clinicians	25
6.1	Motivation	25
6.2	Data	25
6.3	Methods	25
6.3.1	Approach to modelling	27
6.4	Findings and discussion	27
7	Study 2: Partitioning of excess mortality in population-based cancer patient survival studies	30
7.1	Motivation	30

7.2	Data	30
7.3	Methods	31
7.4	Findings and discussion	33
8	Study 3: Temporal trends in mortality from diseases of the circulatory system after treatment for Hodgkin lymphoma	36
8.1	Motivation	36
8.2	Data	36
8.3	Methods	36
8.3.1	Approach to modelling	36
8.3.2	Predicting excess mortality and crude probabilities of death into the future	37
8.4	Findings and discussion	38
9	Study 4: The application of cure models in the presence of competing risks	41
9.1	Motivation	41
9.2	Data	41
9.3	Methods	42
9.3.1	Approach to modelling	43
9.4	Findings and discussion	43
10	Conclusions and discussion	46
10.1	What do we aim to estimate?	46
10.2	Presentation of results	50
10.3	Statistics targeted towards lay audiences	51
11	Future perspectives	52
11.1	Adjusting for comorbidities in studies of cancer patient survival	52
11.2	Assumptions for the expected survival in the cancer-free population	53
12	Acknowledgements	54

List of abbreviations

The following abbreviations are used throughout this thesis:

$S(t)$	All-cause survival
$S^*(t)$	Expected all-cause survival in a population assumed free from the disease under study
$R(t)$	Relative survival
$h(t)$	All-cause hazard
$h_c(t)$	Cause-specific hazard
$h^*(t)$	Expected hazard in a population assumed free from the disease under study
$\lambda(t)$	Excess hazard
$H(t)$	Cumulative all-cause hazard
$H^*(t)$	Cumulative expected hazard
$\Lambda(t)$	Cumulative excess hazard
$F_c(t)$	Cause-specific cumulative distribution function
$Cr_c(t)$	Crude probability of death from the diagnosed cancer
$Cr_o(t)$	Crude probability of death from causes other than the diagnosed cancer
$Cr_{alive,c}(t)$	Crude probability of being alive and predicted to die from the diagnosed cancer
$Cr_{alive,o}(t)$	Crude probability of being alive and predicted to die from causes other than the diagnosed cancer
π	Cure proportion
$P_{btd,c}$	Proportion of personally cured patients
CDR	Cause of Death Register
ICD	International Classification of Diseases
INCA	Information Network for Cancer Care
NPCR	National Prostate Cancer Register
PSA	Prostate Specific Antigen
RCC	Regional Cancer Center
SCR	Swedish Cancer Register
TNM	Tumor, Node, Metastasis
HMD	Human Mortality Database

1 Introduction

In Sweden, the collection of individual-specific data in population-based health registers has been routine for more than half a century. Consequently, register-based research has become an integral part of the Swedish epidemiological research tradition. It is not surprising that continuous efforts are made to nurture this tradition, and to push the limits for how health registers are used in research [1]. As we broaden the focus from classical epidemiological studies, to include also more complex clinical research studies, we need to continue developing appropriate statistical methods to ensure that maximum use is made of the data at hand. We must also realize that the tools we develop for this purpose are typically not self-explanatory, and that considerable effort must be made to communicate new methods and make them accessible and understandable for their target users.

Cancer statistics have a broad range of consumers, ranging from researchers, physicians, patients, patient organisations, policymakers to medical (as well as non-medical) journalists. Although many of the classical methods for producing and presenting statistics of cancer patient survival are useful for aetiological and public health research, they may be less optimal for communication in a clinical setting. For example, the most commonly reported statistics for population-based cancer patient survival, relative survival, is interpreted in the hypothetical scenario where cancer is the only possible cause of death. From the perspective of newly diagnosed patients, who seek to understand the potential impact of the cancer diagnosis on their life expectancy, such statistics may be of little help. Ideally, clinically relevant statistics on prognosis of cancer patients should enhance the information present in the data, be presented in a manner that can aid physicians in risk counselling situations, and facilitate informed decisions about the clinical management of cancer patients. Even though statistical methodology that aims to fulfil this purpose is regularly developed, the uptake of the new statistical theory to applied research in a broad sense is typically slow. The reason for this is most likely multi-factorial but possible explanations include:

- new statistical theory does not always reach audiences beyond the technical journals where the method was first published
- a lack of a clear link between the applied research question and the new statistical theory
- the new methodology is not implemented in statistical software used by applied researchers
- an unwillingness among applied researchers to abandon traditional, and widely accepted statistical methods (old habits are hard to kick!)

This thesis strives to overcome some of these barriers by combining the development of new statistical theory, needed to appropriately answer clinical research questions, with illustrative examples of the methods using real data. Crucial to this endeavor has also been to motivate and disseminate non-standard statistical methods to their target audiences by providing easy access via

implementation of the methods in user-friendly software, teaching short courses, and by publishing educational scientific papers in epidemiological and clinical journals.

2 Aims of the thesis

The overall aim of this research is to develop statistical methods for estimating and modelling cancer patient survival, with particular emphasis on developing and applying new methods for presenting population-based cancer survival statistics in a manner relevant for physicians and patients.

Specifically, the aims were to:

- Appraise statistical methods for competing risks in a population-based setting and to discuss interpretation, assumptions, strengths and weaknesses in relation to other, more traditional statistical methods for population-based cancer patient survival (Study I).
- Extend flexible parametric survival models, adapted for relative survival, in order to estimate treatment-related mortality in the presence of competing risks (Studies II and III).
- Combine the use of flexible parametric cure models with competing risks theory in order to estimate the proportion of patients not bound to die from their diagnosed cancer (Study IV).
- Demonstrate how period analysis and conditional estimates of survival can be applied in combination with the proposed statistical methods to increase the usefulness of the results for newly diagnosed cancer patients (Studies I and IV).
- Implement the developed statistical methodology in user-friendly software to make it readily available to applied users (Studies II and IV).

3 Background

3.1 Cancer

The human body consists of 100 trillion (10^{14}) cells. These are classified into over 200 distinct cell types, where each type has a specialised task to perform. Cancer is the shared name for diseases that can arise in almost any of the different cell types. As such, cancer is not one but many different diseases. The main characterising features of carcinogenesis, i.e., the process that transforms healthy cells into cancer cells are, uncontrolled proliferation (destabilised cell division), dedifferentiation (loss of specialisation), impaired ability to undergo normal apoptosis (programmed cell death), acquired autonomy from other cells (self-sufficient in growth signals and blood supply), and loss of capacity to repair genetic errors [2]. The reprogramming of cells occurs as a consequence of genetic mutations and epigenetic changes in healthy cells. Whilst reprogramming of one single healthy cell into a cancer cell is sufficient to give rise to cancer, several changes to the genetic material of healthy cells are required to initiate and promote a malignant transformation. In most cases, the process in which one single ancestral cancer cell develops into detectable cancer takes many years, although the natural history of cancer varies heavily across different cancer types. Unless the disease is successfully treated or controlled, cancer might spread via the blood and lymphatic system and invade other organs of the body by forming metastases. Treatment options for most cancers include removal of solid tumours by surgery, targeting rapidly dividing cells with chemotherapy, shrinking tumours and destroying cancer cells by radiation, blocking the growth of cancer cells by targeted therapies, or inducing the patient's own immune system by immunotherapy. The aim (curative or palliative) and choice of treatment depends on the type, location and spread of the disease, as well as on the health status of the patient. Metastasized cancer is, for example, generally not amenable to cure, and the treatment is thus primarily palliative.

Cancer is generally regarded as a chronic disease. As such, it is a huge burden on our society, both economically and socially. Annually, more than 50,000 new cases are diagnosed in Sweden, and approximately 20,000 deaths from cancer occur [3]. About 50% of all new cases occur in individuals older than 70 years, and one third of the population will be diagnosed with cancer before their 75th birthday [3]. The number of individuals who live with cancer has increased over time, and is estimated to further increase in the future as many types of cancer can be detected at an earlier stage than before, and treatment regimens are becoming increasingly successful. Moreover, the age structure in Sweden is changing with the population aging [4]. The growth of the elderly population and the increase in the proportion in older age groups, further adds to the cancer burden and resources that are used for cancer care. In addition, the social consequences stretch beyond the individuals who are personally afflicted by cancer since virtually everyone will have an affected family member or friend at some time during their lives.

The known causes of cancer can be divided into heritable factors, life-style factors, reproductive factors, infectious agents, exposure to radiation (ionising and non-ionising ultraviolet radiation),

workplace/household carcinogens and pollution. Lifestyle factors such as smoking, diet, and physical activity are estimated to cause about 70% of all cancers [5]. Smoking alone is believed to be attributable to 15% of all cancers, whereas heritable factors give rise to approximately 5 to 10% of all cases [5].

3.2 The role of a cancer registry in cancer control

The Swedish Cancer Registry was founded in 1958 with the purpose of creating a data base with national coverage that could be used to map the occurrence of cancer, monitor temporal changes in incidence, mortality and survival, facilitate clinical and epidemiological research and make international comparisons possible [6]. This role has remained virtually unchanged until today although organisational changes for oncologic care in Sweden have led to the majority of the registration work and activities related to cancer control now taking place in Regional Cancer Centers. Cancer registries have an important role in cancer control programmes. As defined by the World Health Organisation [7],

"a national cancer control programme is a public health programme designed to reduce the number of cancer cases and deaths and improve quality of life of cancer patients, through the systematic and equitable implementation of evidence-based strategies for prevention, early detection, diagnosis, treatment, and palliation, making the best use of available resources."

Classic components of a cancer control programme include activities related to 1) primary prevention, 2) screening, 3) enabling early diagnosis, 4) optimisation of treatment, 5) rehabilitation, and 6) palliative care [8]. Cancer registries have a unique role in these activities since their wealth of data can be used in research that monitors and evaluates the performance of existing programmes, as well as to carry out needs assessment for future programmes [9, 8, 10]. From an international perspective, no cancer registry is currently involved in each and every component of cancer control. Like in society at large, resources are limited and the goals and tasks in an organisation are balanced against their benefits and their costs. For cancer registries, the implication is that their direct role in cancer control is somewhat fragmented. In a survey of the members and associate members of the International Association of Cancer Registries (IACR), Armstrong reviewed what activities (related to cancer control) that cancer registries worldwide were most frequently involved in. Most contributions were in basic research on cancer prevention (epidemiological research into the causes of cancer), clinical trials of treatment, and the production of incidence statistics. Other, less commonly, reported activities were related to situation analysis, coordination and monitoring of screening programmes, and public education [8]. In practice, many external organisations and research groups also contribute to the various parts of cancer control by using registry data for aetiological research, evaluation of primary and secondary prevention programmes, health care planning and patient care. In such situations the cancer registry does not participate directly in the research and surveillance activities, but serves as the collector and provider of research data.

3.3 Outcome measures used in cancer control

The statistical measures that are typically used in cancer control include incidence, mortality and survival [11]. *Incidence* is the most frequently reported measure of cancer burden. Trends in incidence, i.e., the rate at which new cancer cases arise in a given population, mimics trends in the distribution of risk factors and can therefore be used to monitor the performance of efforts that have been taken to prevent disease occurrence (i.e., activities related to primary prevention).

Mortality reflects death in the population and is the principal outcome measure for evaluating the success of screening programs. Mortality is calculated by dividing the number of deaths from cancer by the total population at risk of dying over some well-defined time period. As such, trends in mortality are not only mimicking trends in cancer survival but also trends in the incidence. The accuracy of mortality statistics is closely linked to the accuracy of the causes of death reported on death certificates [12].

To evaluate changes in the progress of diagnosing and effectively treating cancer patients, estimates of *survival* (or equivalently, mortality restricted to the population of cancer patients) is the most appropriate measure. Calculation of survival requires patient follow-up (from the date of diagnosis until death or censoring) and provides an estimate of the proportion of patients who are still alive as a function of time since their diagnosis. Survival is the cornerstone of all studies included in this thesis and we will return to the topic of estimation and interpretation of survival in the coming chapters.

In addition to incidence, mortality and survival, other useful measures include prevalence, and person-years of life lost due to cancer. Common for all measures introduced here is that none take into account the morbidity that patients might suffer from their disease. Quality of life adjusted outcome measures are needed to fully account for all components of cancer control but a detailed review of this aspect of patient care is beyond the scope of this thesis.

3.4 Communicating and producing statistics relevant for clinicians and patients

Whilst all of the above mentioned outcome measures have an important role in cancer control, their usefulness is limited for an individual patient, whose primary interest might be in understanding how the disease, and the treatment of it, will affect her life. Psychosocial aspects, such as feelings of guilt, anxiety and fear, have long been recognised as important issues in cancer patient care, but patient reactions to cancer are not only highly variable between individuals, but also during different phases of disease. The coping strategy for one patient might be to actively seek information about the disease, in order to gain an understanding of the severity of the situation and the expected course of treatment, whereas another patient might prefer taking a more passive role. In many countries, including Sweden, the general idea in patient education and rehabilitation activities is that an active role, that involves confronting facts about the disease and, if possible, participating in decisions about the management and care, is more beneficial for the patients than a passive role [13]. However, adjusting to a life as a cancer patient is a long process, that often starts in a state

of shock and includes disease reaction, anxiety management, reality adjustment, and eventually finding a way back to everyday life. It is important to respect that not all patients are ready to take an active role during all phases of this process.

Risk counselling lays the foundation for the information and tools that patients require in order to become active participants. It involves disseminating the most recent information about cancer risk and prognosis from the learning continuum of research, via health care professionals, to cancer patients and their families. The general idea behind any form of risk counselling is to provide patients with a sense of coherence and encourage informed participation in medical care. The success of risk counselling is, nevertheless, highly dependent on the communication skills and information literacy and numeracy of both the counsellor and counselee, as well as the understanding of the source data that forms the basis for the information [14].

This thesis is concerned with producing statistics that are useful to predict and summarise the most likely outcome of the disease. There is no standardised way in which such information is communicated between clinicians and patients today, and there is not even agreement that information that stems from grouped patient data should be presented and discussed routinely with patients. Notwithstanding this, it is recommended that an explicit wish from the patient to discuss the anticipated prognosis, and the potential risks and benefits of the treatment, in terms of actual percentages should be respected [15]. At first glance, estimates of survival seem to be the most suitable outcome measure to attempt to answer such questions, and in practice, the five-year survival probability is often reported and quoted as a measure of prognosis. There are however, several pitfalls with a naive use of survival proportions in risk counselling between physician and patient which leave room for misuse and misinterpretation.

Firstly, the estimated proportion of patients who survive an arbitrary number of years after their initial cancer diagnosis is typically calculated by the use of explicit information about the recorded cause of death, or via a relative survival framework. Both these approaches aim to estimate net survival. As such, the correct interpretation of, for example, the five-year cancer specific survival proportion is: *"the proportion of patients who survived five years after their diagnosis under the assumption that only the diagnosed cancer could kill them."* The subtle assumption that the patients are assumed immune from competing causes of death is often not fully understood and therefore sometimes over-looked in the communication of survival statistics in the context of cancer prognosis. Secondly, communicating the relevant time frame for which survival statistics are valid is crucial. A statement like *"The survival of patients with similar disease and patient characteristics as you has been reported to be around 70%"* can be grossly misleading since it entirely lacks references to time. Does it suggest that 70% will not die from the cancer within a specific time window after diagnosis (e.g., 5 years), or perhaps that 70% of the patients will never die from the diagnosed cancer? What if the patient has already survived one year since the diagnosis? How does that fact alter the above statement? Moreover, from what cohort of patients was the stated survival proportion calculated? Was it based on patients diagnosed many years in the past? If so, how are changes in the diagnosis and treatment that might have taken place over time expected to change the prediction for patients

diagnosed today?

Thirdly, the population-based data from which grouped estimates of survival often come, is inherently observational and must always be interpreted in this context. Communicating uncertainty in predictions and explaining that there is no such thing as being able to predict the exact course of disease for a specific individual is a challenging task for any provider and communicator of statistics targeted to patients or to the general public.

Lastly, even the most carefully calculated and interpreted estimates of survival do not account for other aspects of treatment and rehabilitation, such as quality of life issues that might be of equal, or even greater concern to the patients than the actual prospect of surviving five more years.

The bottom line is that all attempts to communicate medical information involves translating, often high-level, information into meaningful messages for different types of audiences. This requires:

1. a good understanding of what type of information is expected from the counselee.
2. easy-access to systematically structured and up-to-date disease-specific data.
3. standardized communication materials to avoid misinterpretation resulting from inconsistent presentation.
4. comprehensive disease-specific tools that facilitate patient comprehension and that can be adapted for individual levels of numeracy, literacy and susceptibility to framing effects (i.e., how small changes in the format of numerical information may dramatically alter the perception of the presented information [16]).

Recommendations that meet the above requirements and that aim to optimize risk communication between doctor and patient are, however, not always straight forward to implement in practice. For example, while there is a consensus that the main care provider has the greatest insight into the clinical management of the individual patient, factors that include lack of easy-access to appropriate data and tools for satisfactorily explaining the data to the patient, limited face-to-face time for patient consultations, and uncertainty whether the patient actually is ready for, or even wants to take an active role in her management, adds complexity to risk communication. In addition, individuals diagnosed with severe illnesses such as cancer have been reported to also seek information about their disease outside the health care system. In the end of 2011 in Sweden, 89% of all Swedish households were equipped with a computer that had access to the internet, 71% of all Swedes were reported to access the internet on a daily basis, and 23% of the users looked for health information online at least monthly according to a report issued by the World Internet Project [17]. Among individuals diagnosed with cancer in Sweden, the proportion of patients that use the internet to access information about their disease has increased markedly in the past 15 years, going from 16% in 1998 to 61% in 2008 [18, 19]. Swedish cancer patients have also been found to pay closer attention to media reports about cancer than non-cancer patients [18] which puts high demands not only on health care professionals, but also researchers, medical reporters, support groups and other providers of health-related information to the public.

4 Material

4.1 Cancer registration in Sweden

Established in 1958, the Swedish Cancer Register (SCR) is one of the world's oldest cancer registers. The SCR is maintained by the National Board of Health and Welfare and encompasses all individuals with an official residency in Sweden. Notification of all malignant tumours (as well as certain benign tumours) is statutory according to the Health Data Registers Act [20], and reports of incident tumours are typically submitted from at least two sources (physicians, pathologists or cytologists) [21]. National cancer registration in Sweden is based on data collected regionally via six Regional Cancer Centers (RCC). Each RCC is situated in one out of six health care regions in Sweden, and the notification of new cancer cases are continuously recorded via the RCCs. The RCCs not only enters the reported tumours into a computerised database, but also perform coding and logical checks, including investigation of cases that lack complete clinical, histopathological or cytological reports. In October each year, the data from the previous calendar year is delivered from each RCC to the National Board of Health and Welfare, and included the SCR. The forwarded data consist primarily of information on diagnoses codes, histological subtype, stage (available since 2004), means of detection, and where the cancer was diagnosed.

In addition to preparing and delivering regional cancer data to the National Board of Health and Welfare, the RCCs also host 28 cancer quality registers. The quality registers are decentralised health registers that are administered via a common IT-platform, INCA (Information Network for Cancer Care), which is maintained jointly by the RCCs. Each quality register has been initiated by physicians and other health care professionals, and is used for monitoring the quality of the care and management of cancer patients in Sweden. The scope of information related to patient (host), tumour and treatment that is reported to, and recorded in, the quality registers is more detailed than that of the national SCR. As such, the quality registers are often used in register-based epidemiological studies, as well as a starting point for clinical research. A summary of the data flow and organisational structure of Swedish cancer registration is summarised in Figure 1.

4.2 Data material used in this thesis

4.2.1 Cancer data

In this thesis, data from the SCR was used in studies II (breast cancer), III (Hodgkin lymphoma) and IV (melanoma, colon cancer and acute myeloid leukemia). For study I, data from the National Prostate Cancer Register (NPCR) was used. The NPCR is a quality register that was established in the South-East health care region of Sweden in 1987, and which reached national coverage in 1998. The NPCR records all new cases of prostatic adenocarcinoma and is regularly updated against the SCR. In terms of completeness, the NPCR encompasses more than 96% of all prostate cancer reported to the SCR [22]. In addition to basic demographic data of the patient, the NPCR also contain detailed clinical data, such as, stage according to the TNM-classification system, tumour

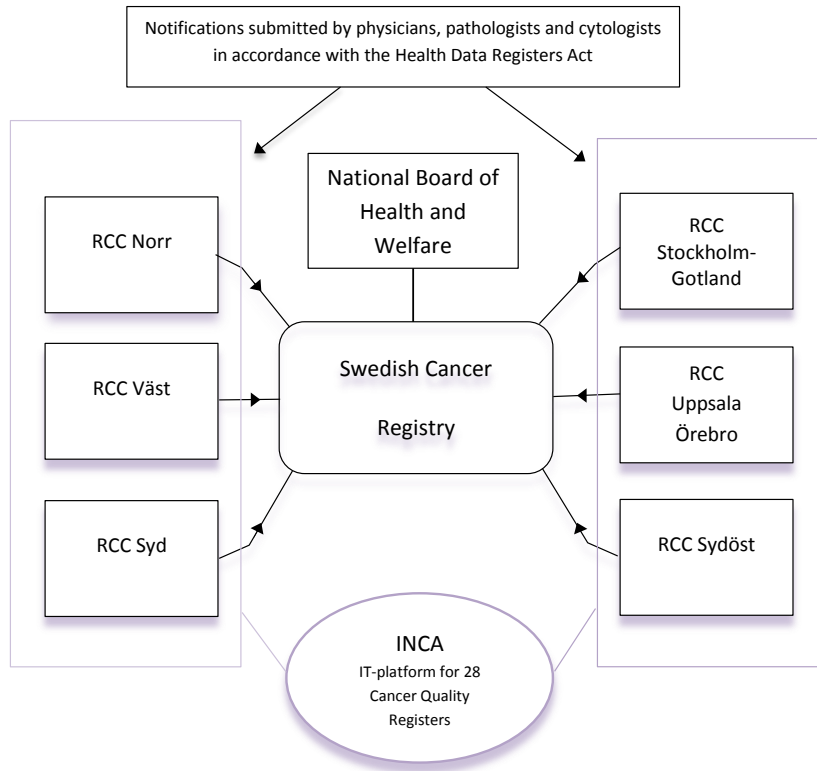


Figure 1 – Organisational structure of cancer registration in Sweden

grade according to the Gleason grading system, serum Prostate Specific Antigen (PSA) level at the time of diagnosis, as well as intended treatment.

4.2.2 Death data

Both the SCR and the NPCR are regularly linked to the Swedish Cause of Death Register (CDR) to retrieve information on dates of death (if applicable) that enable patient follow-up. The CDR was initiated in 1961 and includes all deceased individuals who were residents in Sweden at the time of their death. Stillborn children, temporary residents, and Swedes who had officially emigrated from Sweden at the time they died are not included in the register. However, this is usually not a problem in cancer patient survival studies since the Swedish Cancer Registry have merged the SCR to data that contains information about all recorded emigrations from Sweden. The emigration dates are used in our analyses to censor individuals who emigrate during follow-up.

The CDR also includes information about the underlying and contributing causes of death, classified according to the version of the International Classification of Diseases (ICD) in force at the time. All studies in this thesis used dates of death linked from the CDR. In addition, Study II and III used explicit information about the underlying cause of death of deceased individuals in order to identify patients whose death was attributable to diseases of the circulatory system.

4.2.3 Population Statistics

Sweden has an over 300 year long history of keeping and continually updating a population register containing information about the actual number of people living in the country. Initially, registration was made possible via church records in each parish, but from 1749 population statistics were systematically collected and summarized by *Tabellverket*. However, it was not until after a major re-organisation of the compilation of population statistics in 1858, when Statistics Sweden was founded, that the quality of the statistics became satisfactory. As of 1860, the data on population and death counts is thought to have been nearly 100% complete [23]. Since then, Statistics Sweden has remained the administrative agency responsible for producing and publishing official Swedish statistics for public information, planning and research purposes. Today, the population statistics are compiled using data from population records at the Swedish Tax Agency.

In this thesis, population life tables, stratified on age, sex and calendar year, were used in all four studies. The life tables were, nevertheless, not obtained directly from Statistics Sweden but from the Human Mortality Databases (HMD) project. The HMD is a collaborative project that was launched in 2002 in order to provide open, international access to detailed mortality and population data to researchers, journalists, policy analysts among others. It is a joint project that primarily involves research teams in the Department of Demography at the University of California, Berkeley, USA, and at the Max Planck Institute for Demographic Research, Germany. The Swedish raw data is submitted to the HMD by Statistics Sweden and adjusted and processed according to a common methods protocol prior to being converted to the life-tables that are available for public use [24].

4.3 A note on record linkage

Data from the population and health registers were linked using the Swedish personal identity number that is assigned to every person who is born in Sweden, as well as immigrants who intend to stay in Sweden for at least one year. The personal identity number was first introduced in Sweden in 1947, and has since become a vital component of Swedish health care administration by enabling tracing of patients and their medical records, and register linkages for research purposes [25]. The personal identity number consists of three parts; the date of birth, a three-digit birth number and a check digit and makes it possible to trace virtually any patient through, not only the national health registers, but also other registers that contain population-statistics, such as migration, education, taxation, and income statistics. Medical research projects that require record linkage must be reviewed and granted by an Ethics review-board, and in practice, the linkage is done at the National Board of Health and Welfare or Statistics Sweden where the personal identity number is replaced by a serial number in order to make the individual records anonymized before the data is released to researchers.

5 Statistical methods

5.1 Relative survival and excess mortality

All statistical methods used in this thesis build upon the theory of relative survival. The concept of relative survival was introduced for the first time in 1942 [26] and has since then become the method of choice for estimating population-based cancer patient survival [11]. Relative survival provides a measure of excess mortality associated with cancer, i.e., the mortality rate above and beyond what is expected in a comparable group of individuals free from the cancer in question. It is calculated without using explicit cause of death information, by contrasting the all-cause survival experienced by the patients, to the expected survival in a comparable group of individuals, free from the cancer in question. Formally relative survival, $R(t)$ at time t after diagnosis can be written as

$$R(t) = \frac{S(t)}{S^*(t)}, \quad (1)$$

where $S(t)$ is the all-cause survival among the patients and $S^*(t)$ is the expected (all-cause) survival in the comparison group, free from the studied cancer. The excess mortality, denoted by $\lambda(t)$ at time t is

$$\lambda(t) = h(t) - h^*(t), \quad (2)$$

where $h(t)$ is the all-cause mortality rate experienced by the patients and $h^*(t)$ the corresponding expected mortality rate. The expected survival (and mortality) is typically obtained from the National Bureau of statistics and stratified on characteristics such as sex, age and calendar year. Relative survival *aims* to provide an estimate of net survival. Net survival is interpreted as the survival experience of the patients in the absence of competing causes of death. For example, a five-year net survival of 0.8 provides an estimate of the proportion of patients who would survive at least 5 years after their diagnosis of cancer, 80% in this example, if it was possible to remove all causes of death except the cancer in question. Although there is occasionally a mistake belief that relative survival and net survival represent the same quantity, the interpretation of relative survival as net survival is only valid under the following assumptions,

1. The time to death from the cancer in question is conditionally independent of the time to death from other causes. By conditional independence we mean that there are no factors that influence both cancer and non-cancer mortality other than those factors that have been controlled for in the estimation via stratification or regression modelling. This is commonly referred to as the *independence assumption*.
2. The cancer patients are exchangeable to the comparison population that gave rise to the expected mortality rates (typically conditional demographic covariates like age, sex and calendar year).
3. An appropriate estimator is used.

Due to the fact that the cancer and non-cancer mortality typically share the influence of the same demographic covariates, it has been shown that the most commonly used estimators for calculating a single summary measure of the life-table estimates of relative survival do, in fact, yield biased estimates of net survival [27]. The reason for this is that a single (averaged) expected mortality rate is applied to a group of heterogeneous individuals. However, adjusting the relative survival estimates for age, e.g., by stratification or standardisation, reduces the bias in all classical estimators [28]. The recently proposed Pohar-Perme method, on the other hand, yields unbiased estimates of net survival for a group of cancer patients as a whole, in a setting that does not require stratification or regression modelling. The Pohar-Perme method is therefore anticipated to become the preferred method for reporting cancer patient survival statistics in the future in situations where one single summary measure of cancer patient survival is of interest, e.g., in cancer registry reports [29, 30, 31].

In this thesis we will use multivariable regression to model relative survival (excess mortality). Individual-level modelling of the excess mortality removes, to a certain extent, the problem raised above since it appropriately incorporates the demographic covariates in question in the estimation [27, 28]. However, any regression model is dependent upon the validity of all additional assumptions that are made in the model. For example, if there are additional covariates not included in the model that share influence on cancer and non-cancer mortality, the model-based relative survival estimates will not be estimates of net survival.

5.1.1 Modelling excess mortality using flexible parametric survival models

In survival analysis, regression models are usually fitted on the (log) mortality scale, as opposed to the survival scale. This is also the case in relative survival analysis, and a range of regression models for excess mortality have been proposed. To date, models where the baseline excess hazard function is assumed to take a parametric form, and where covariates are assumed to act multiplicatively on the excess hazard, have been used most commonly in applied research [32, 33, 34, 35, 36, 37]. Non-parametric [38, 39] and fully additive regression models [40, 41, 42] have also been developed to model excess mortality but have been used less often in practise in population-based cancer patient survival analysis.

In this thesis, flexible parametric survival models were used. Flexible parametric models are fitted on the log cumulative excess hazard (mortality) scale and use restricted cubic spline functions to explicitly model the (log cumulative) baseline excess mortality function [37]. That is, on the cumulative hazard scale the all-cause cumulative mortality, $H(t)$, at time t is

$$H(t) = H^*(t) + \Lambda(t), \quad (3)$$

where $H^*(t)$ is the cumulative expected all-cause mortality in the general population (matched on sex, age and calendar year), and $\Lambda(t)$ is the cumulative excess mortality. The flexible parametric survival model assumes that $H^*(t)$ is a known quantity and that $\Lambda(t)$ is the random component subject to regression modelling. By assuming that the cumulative excess hazard is a multiplicative

function of the covariates, \mathbf{x} , and that the effects are proportional with respect to the underlying time scale, a model for $\Lambda(t)$ can be written as

$$\ln(\Lambda(t; \mathbf{x})) = s(\ln(t); \boldsymbol{\gamma}_0) + \mathbf{x}^T \boldsymbol{\beta}. \quad (4)$$

Here $\Lambda(t)$ is represented by restricted cubic splines for $\ln(t)$, $s(\ln(t); \boldsymbol{\gamma}_0)$, characterised by the vector of parameters associated with the basis functions for the spline, $\boldsymbol{\gamma}_0$, and the effects of covariates, \mathbf{x} , which are given by $\boldsymbol{\beta}$. The restricted cubic spline function, $s(\ln(t); \boldsymbol{\gamma}_0)$, is defined as,

$$s(\ln(t); \boldsymbol{\gamma}_0) = \gamma_{00} + \gamma_{01}v_1(x) + \gamma_{02}v_2(x) + \dots + \gamma_{0K-1}v_{K-1}(x), \quad (5)$$

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

$$v_j(x) = \begin{cases} x, & \text{if } j = 1 \\ (x - k_j)_+^3 - \lambda_j(x - k_{\min})_+^3 - (1 - \lambda_j)(x - k_{\max})_+^3, & \text{if } j = 2, \dots, K - 1. \end{cases} \quad (6)$$

Under this definition, $u_+ = u$ if $u > 0$ and $u_+ = 0$ if $u \leq 0$, k_{\min} is the position of the first knot, k_{\max} the position of the last knot, and $\lambda_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}$.

The proportional excess hazards model in (4) can easily be extended to handle time-dependent covariate effects by including interaction terms between the time-dependent covariates and a new set of restricted cubic spline terms. Modelling of time-dependent effects generally require fewer knots than the baseline effects [37] so for each time-dependent effect, x_i , a new configuration of the knots may be chosen. This gives,

$$\ln(\Lambda(t; \mathbf{x})) = s(\ln(t); \boldsymbol{\gamma}_0) + \mathbf{x}^T \boldsymbol{\beta} + \sum_{i=1}^D s(\ln(t); \boldsymbol{\gamma}_i) \mathbf{x}_i \quad (7)$$

where D is the number of time-dependent covariate effects and $s(\ln(t); \boldsymbol{\gamma}_i)$ is the spline function for the i^{th} time-dependent effect. Note that for each of the D time-dependent effects represented by x_i in the model above are typically a subset of \mathbf{x} .

The model parameters are estimated by maximising the log likelihood function derived by Nelson et al [37]. The likelihood contribution for each subject is

$$\ln L_i = d_i \ln \left[h^*(t_i) + \frac{1}{t_i} \frac{d}{dx_i} (s(x_i; \boldsymbol{\gamma}_0)) \exp(\eta_i) \right] - \exp(\eta_i)$$

where $x_i = \ln(t_i)$, and η_i is the linear predictor corresponding to individual i in the flexible parametric model as expressed in (4). The extension to left truncated data can easily be made by adding the term $\exp(\eta_{0,i})$ to the above likelihood contribution for each subject. In this setting, $\exp(\eta_{0,i})$ corresponds to individual i 's linear predictor at her entry time, $t_{0,i}$. That is,

$$\exp(\eta_{0,i}) = s(\ln(t_{0,i}); \boldsymbol{\gamma}_0) + \mathbf{x}^T \boldsymbol{\beta}.$$

The maximisation of the log likelihood function, $\ln L = \sum_i^n \ln L_i$ was done using the Newton-Raphson algorithm [43].

5.2 Period analysis

A potential problem with survival estimates based on patients diagnosed five or ten years back in time is that they might no longer be valid for newly diagnosed patients. The introduction of new effective treatments could, for example, alter the conditions and prospect for recovery and rehabilitation for the patients who receive it, compared to those who previously did not. The survival estimates calculated from the patients who were diagnosed and treated prior to the introduction of the more effective treatment would paint an overly pessimistic picture of the future for patients diagnosed in more recent years. As a means to predict cancer patient survival in a manner that improves the timeliness of survival, estimates *period analysis* was introduced for cancer patient survival by Brenner and Gefeller about 15 years ago [44].

The conceptual idea behind period analysis is illustrated here using a mock sample of 4 cancer patients diagnosed between 1996 and 2003 and followed for death or censoring until the end of 2005. In Figure 2 solid lines have been drawn to represent the length of follow-up of each of the four patients. In a classical complete cohort analysis, subject 1 would enter the cohort at the date of her diagnosis in 2003, and be followed through time until the date of death or censoring (due to loss from follow-up or administrative censoring which occurred in 2005). Thus, subject 1 would contribute to the riskset in calculations of survival up to 2 years after her diagnosis. Similarly, subject 2 would contribute to the riskset up to 4 years after the diagnosis etc. In a period analysis,

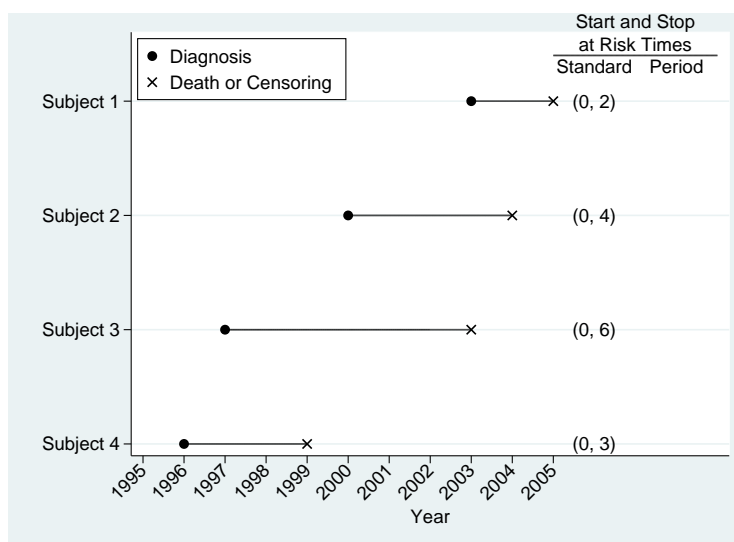


Figure 2 – Cohort of 4 cancer patients and their contribution to a survival analysis

only a subset of each subjects follow-up is incorporated into the survival calculations. Which part is included depends on when the patient was diagnosed with cancer, and on the *period window*. Since period analysis aims to improve the timeliness of survival estimates, a period window that

includes only the most recent years for which data are available is typically chosen. In Figure 3, the years 2002 through 2005 define the period of interest, and only follow-up within this window is incorporated in the survival calculations. For subject 1, this corresponds to the full follow-up, i.e., the same amount of information is included in the period analysis as in a complete cohort analysis. For subject 2 the survival time between the date of diagnosis and year 2002 is ignored, and subject 2's contribution to the riskset is thereby restricted to the survival calculation beyond the 2 first years after the diagnosis. Subject 4, who exited the study prior to the period of interest is entirely left out from the analysis. Thus, only the most recent data are used for estimation of short-term survival, i.e., the time during patient follow-up when temporal improvements in cancer patient survival are typically observed [45]. The timeliness achieved by using only recent data for estimation of short-term survival also carries over to estimates of long-term survival, even though historical data is required to estimate of the latter.

In practice, period estimates are obtained by left-truncating all observations at the beginning of the period window and by right censoring them the end [44]. As such, period estimates do not come from a well-defined cohort of patients but from a *synthetic cohort* with varying degree of inclusion in the analysis. Empirical evaluations of period analysis have demonstrated a very good agreement

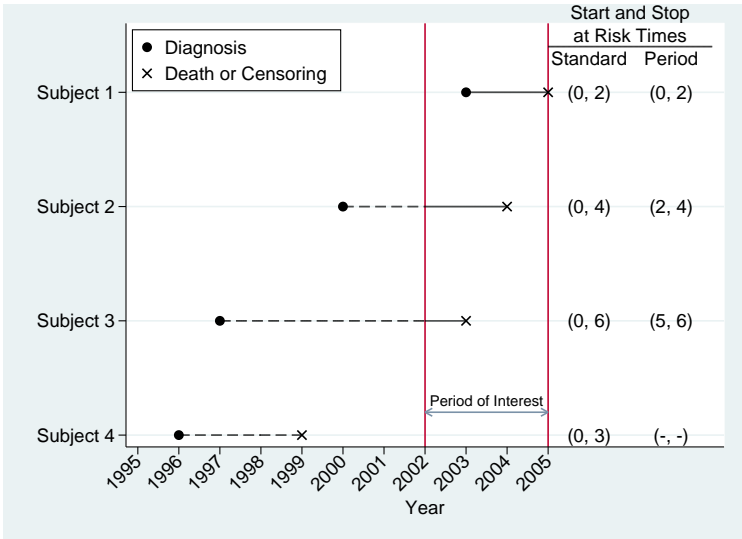


Figure 3 – Synthetic cohort based on 4 cancer patients and their contribution to a period survival analysis

between survival estimates projected using the Brenner approach and what actually happened to the patients once sufficient follow-up was available to calculate traditional cohort estimated of survival [46, 47, 48]. A period approach to estimation was applied in study I of this thesis.

5.3 Statistical methods for competing risks

Under the independence assumption, relative survival aims to estimate net survival¹, i.e., survival in the hypothetical situation where deaths from other causes have been removed. Statistical methods for competing risks, on the other hand, are used to calculate so called crude survival probabilities². These are interpreted as cancer-specific survival probabilities in the situation where other causes of death also exist. From a risk counselling perspective, estimates of net survival may not appropriately describe the most likely course of the disease since they, by definition, will underestimate the actual cancer-specific survival. The distinction between net and crude survival is illustrated in Figure 4 using data from the Swedish NPCR to summarise prostate cancer survival within the first 10 years after diagnosis, among men diagnosed between 2005 and 2009. First we note that the complement of

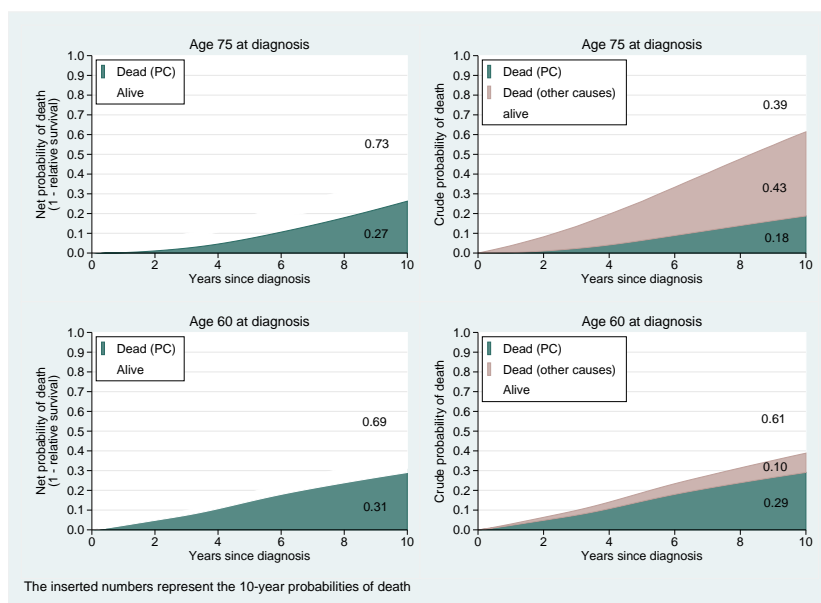


Figure 4 – Net and crude probabilities of death in Swedish men diagnosed with prostate cancer between 2005 and 2009.

the net and crude survival is drawn (i.e., one minus the survival probability), as is tradition in most competing risks applications. The shaded areas in the graphs on the left-hand side represent the net probabilities of prostate cancer death as a function of years since diagnosis. For a 75-year-old man diagnosed with prostate cancer, the probability of dying of prostate cancer within 10 years is 0.27 in a world where it is not possible to die of other causes. When acknowledging the existence of competing causes of death, we see that the crude probability that such a man will die of prostate cancer during the same time interval is only 0.18 (top right hand graph). In the same panel we also see that the probability of dying of a cause other than prostate cancer within 10 years is 0.43, and the probability of still being alive after 10 years is 0.39. The difference between crude and net probabilities is not as great for the 60-year-old men since the probability of death due to causes other

¹Net survival is sometimes referred to as *marginal survival* [49] or *partial crude survival* [50] in the competing risks literature.

²Also referred to as *cause-specific cumulative incidence* [49] or *absolute survival* [51].

than cancer is not as high. However, the two constructs are conceptually different and the choice between net and crude estimates of cancer patient survival depends on the the research question that is to be investigated. In the discussion section of this thesis I will outline some general rules of thumb that may be applied to assess which of the two measures of survival is most appropriate for a given research question, and what quantities can be estimated from the available data.

On a more theoretical note, the crude probability of dying from cancer, denoted $Cr_c(t)$, is commonly expressed by the cause-specific cumulative distribution function, defined to be $F_c(t) = P(T \leq t, \delta = c)$, where δ is an indicator that tells that death from cancer, c , is the event of interest [50]. $F_c(t)$ can be calculated from the cause-specific hazard rates, $h_c(t)$, and the all-cause survival, $S(t)$, via,

$$F_c(t) = Cr_c(t) = \int_0^t S(u)h_c(u)du, \quad (8)$$

where

$$S(t) = P(T > t) = \exp \left\{ - \int_0^t h(u)du \right\} \quad (9)$$

and

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (10)$$

The instantaneous rate at which death due to cause c occurs, $h_c(t)$, is defined to be:

$$h_c(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t, \delta = c | T \geq t)}{\Delta t}. \quad (11)$$

In a relative survival setting, estimates of the all-cause survival, $S(t)$ and cancer-specific hazard, $h_c(t)$, are replaced by their relative survival counterparts. This involves replacing $S(t)$ with $S^*(t)R(t)$ and $h_c(t)$ with $\lambda(t)$, respectively [52]. Hence, the crude probability of death from the cancer in question can now be re-expressed as

$$Cr_c(t) = \int_0^t S^*(u)R(u)\lambda(u)du. \quad (12)$$

Similarly, the crude probability of death due to causes other than cancer, $Cr_o(t)$, is given by

$$Cr_o(t) = \int_0^t S^*(u)R(u)h^*(u)du. \quad (13)$$

5.3.1 Numerical integration and variance approximation

Expressions (12) and (13) were approximated numerically using the 'brute force' method proposed earlier by Carstensen [53]. The time scale (time since diagnosis) was split into 1000 intervals of equal length, and the predicted, interval-specific values of the integrand, $\hat{f}(t_j | \mathbf{x}) = S^*(t_j | \mathbf{x})\hat{R}(t_j | \mathbf{x})\hat{\lambda}(t_j | \mathbf{x})$ were calculated. The crude probabilities of death were then obtained by summing the interval-specific integrands over the 1000 intervals as follows,

$$\begin{bmatrix} \hat{C}r_c(t_1) \\ \hat{C}r_c(t_2) \\ \hat{C}r_c(t_3) \\ \dots \\ \hat{C}r_c(t_{999}) \\ \hat{C}r_c(t_{1000}) \end{bmatrix} = l \times \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 1 & \dots & 1 & 0 \\ 1 & 1 & 1 & \dots & 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{f}(t_1) \\ \hat{f}(t_2) \\ \hat{f}(t_3) \\ \dots \\ \hat{f}(t_{999}) \\ \hat{f}(t_{1000}) \end{bmatrix} = \mathbf{L} \begin{bmatrix} \hat{f}(t_1) \\ \hat{f}(t_2) \\ \hat{f}(t_3) \\ \dots \\ \hat{f}(t_{999}) \\ \hat{f}(t_{1000}) \end{bmatrix}, \quad (14)$$

where l is the interval length, i.e., $\frac{10}{1000}$. $\hat{f}(t_j | \mathbf{x})$ is a non-linear function of regression model parameters and its variance-covariance matrix was obtained using the delta method. In the setting of this study, the variance-covariance matrix of the vector $\hat{f}(t | \mathbf{x})$ is,

$$\text{Var}(\hat{f}(t | \mathbf{x})) = \mathbf{G}\hat{\mathbf{V}}\mathbf{G}', \quad (15)$$

where \mathbf{G} is a matrix of observation-specific derivatives for each parameter in the model, and $\hat{\mathbf{V}}$ is the estimated variance matrix for the model parameters. Similarly, it follows that the variance-covariance matrix of the vector $\hat{C}r_c(t)$ is,

$$\text{Var}(\hat{C}r_c(t)) = \mathbf{L}\mathbf{G}\hat{\mathbf{V}}\mathbf{G}'\mathbf{L}'. \quad (16)$$

\mathbf{G} were approximated numerically using the Stata command `predictnl` [43].

This approach to estimate crude probabilities of death from cancer was the main focus of study I. In study II and III competing risks theory was combined with a novel approach for partitioning excess mortality into component parts, and in study IV competing risks theory was used in combination with a recently developed method for estimating and modelling statistical cure in a population-based setting [54].

5.4 Cure

Cure is a summary measure used for quantifying improvements in cancer care and prognosis. Whilst there does not exist one single unambiguous definition of cure in the context of cancer patient survival, three broad definitions of conceptually different constructs have emerged from the field of clinical oncology [55].

- *Clinical cure* or *medical cure* refers to the situation where the treatment has eradicated all signs and symptoms of a patient's disease, and where the cancer in question will not return. Even in situations when clinical cure is the goal of the treatment, it is intrinsically hard to objectively measure if an individual patient has reached this goal. In terms of the clinical management of the patient, there is little practical benefit from being able to separate clinically cured patients from asymptomatic long-term survivors who live their lives with no evidence of disease. To quote Robert J. Mayer, MD, a senior cancer researcher at Dana-Farber and physician at Dana-Farber/Brigham and Women's Cancer Center, "*Cure is a word that I don't use a great deal; it is promising something that may or may not be possible. Instead I say to patients, You will be alive and well, and in 20 years we'll look back at this and have a chuckle.*" [56]
- In a relative survival framework, *Statistical cure* occurs from the point in time (after diagnosis) when cancer patients die at the same rate as a comparable group of individuals, free from the cancer in question. In contrast to clinical cure, statistical cure is defined on a group level and is relatively straight forward to estimate. As such, the first cure model was introduced by Boag in 1949 [57]. Informally, statistical cure can be estimated by defining an asymptote to a net survival curve. Figure 5 demonstrates this idea and illustrates how male patients diagnosed with colon cancer can be partitioned into a statistically cured group and a group bound to die from their cancer. Because the cured proportion is identified in a net survival framework, the interpretation of this measure is limited to the hypothetical situation where competing causes of death are assumed not to exist.
- *Personal cure* is a quantity that is not reported often in practise, but which corresponds to the proportion of cancer patients who are likely to die from a cause other than their cancer. The terminology is somewhat misleading since personal cure is, in fact, also defined on a group level. Moreover, personally cured patients do not necessarily die free from signs and symptoms of their cancer. In contrast, the patients may die *with* their cancer. Terminology apart, this measure is of potential interest to cancer patients who want to understand the likely course of their disease since its estimation requires consideration of competing risks. Figure 6 use the same cohort of male patients as in Figure 5 to illustrate the various groups into which patients may be classified after they are diagnosed with cancer. At the time of diagnosis all patients are assumed to be alive. As time progresses some patients will die and,

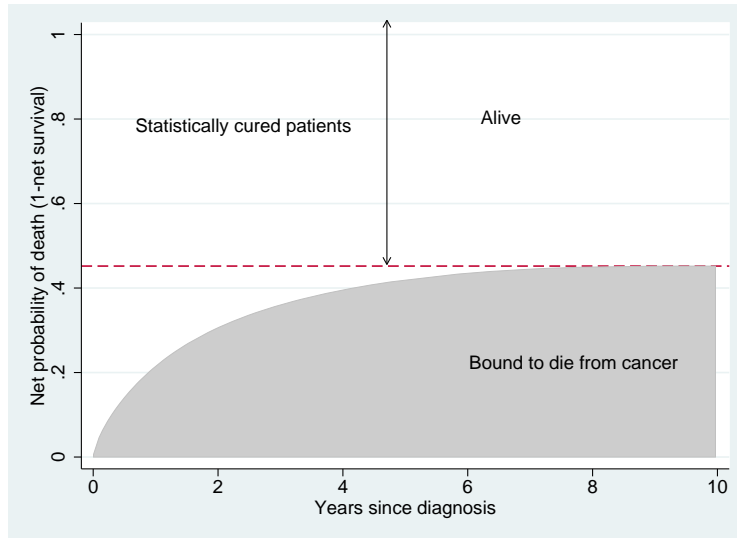


Figure 5 – Conceptual illustration of statistical cure in 75-year old Swedish males diagnosed with colon cancer in 2005

at a given point in time during follow-up, the probability that the deaths that have occurred were caused by the diagnosed cancer versus other causes can be calculated. The proportion of personally cured patients is obtained by estimating an asymptote to the crude survival curve. It will, by definition, be lower than the proportion of statistically cured patients.

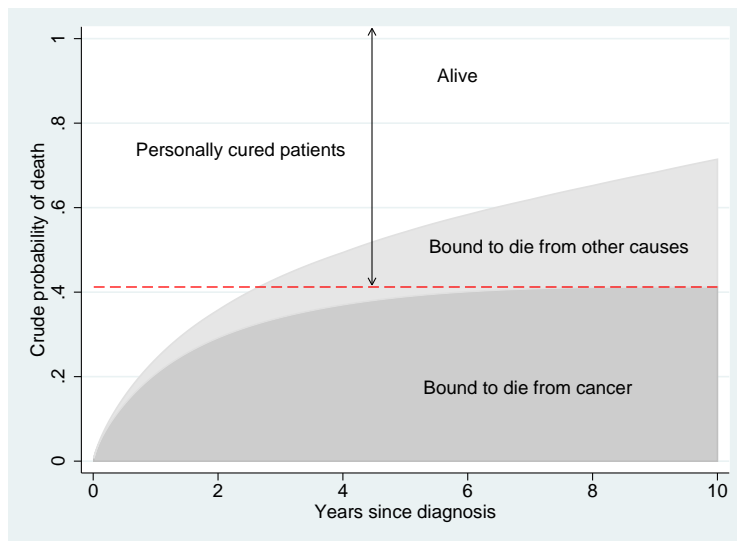


Figure 6 – Conceptual illustration of personal cure in 75-year old Swedish males diagnosed with colon cancer in 2005

5.4.1 Modelling statistical cure

Several models for estimating statistical cure using population-based data have been proposed by, for example, De Angelis [58], Verdecchia [59], Yu [60], Lambert [61] and Andersson [54]. Apart from providing an estimate of the proportion of statistically cured patients, cure models also estimate the

survival function of those that are uncured. Cure models can generally be divided into two classes of models, mixture models and non-mixture models.

In mixture models, the all-cause survival function, $S(t)$ can be written as,

$$S(t) = S^*(t)(\pi + (1 - \pi)S_u(t)), \quad (17)$$

where π is the proportion of patients that will be statistically cured and $S_u(t)$ is the cancer-specific survival function for the uncured patients [59]. Thus, mixture models build upon the idea that patients belong to either a cured group (for which the survival function is assumed to equal that of a comparable disease-free population), or to an uncured group (whose overall survival function can be described by $S^*(t)S_u(t)$). Mixture models primarily differ with respect to the parametric distribution that is used to appropriately model, and capture the shape of the cancer-specific survival function for the uncured group.

In non-mixture models, the all-cause survival function is instead written as,

$$S(t) = S^*(t)\pi^{F_z(t)}. \quad (18)$$

Again, π denotes the cure proportion, whereas $F_z(t)$ denotes some cumulative distribution function, implying that $F_z(t) = 1 - S_Z(t)$, where $S_Z(t)$ is a survival function [61]. $F_z(t)$ can be estimated parametrically, and in practice, the Weibull distribution is often used for this purpose. In contrast to mixture cure models, the original non-mixture models have a biological motivation that stem from what is known about tumor growth [62]. The idea is that treated cancer patients may be left with cancer cells that are capable of forming a new tumour mass, and thereby lead to relapse. Cured patients are assumed to have no such cells left after treatment, whereas for uncured patients, $F_Z(t)$ represents the distribution function that characterises the time it takes for cancer cells with proliferate potential to develop into a detectable cancer. This biological motivation for non-mixture cure models is, however, not relevant in a relative survival setting but the mathematical results can still be applied to estimate an asymptote representative of the cured proportion (as long as statistical cure is believed to be a reasonable assumption).

Equation (18) can, nevertheless, be rewritten as,

$$S(t) = S^*(t) \left(\pi + (1 - \pi) \left(\frac{\pi^{F_Z(t)} - \pi}{1 - \pi} \right) \right). \quad (19)$$

Under this form, the survival function of the non-mixture cure model takes the form of a mixture cure model. Using equation (19) it becomes clear that the survival function of the uncured patients can be obtained via $\frac{\pi^{F_Z(t)} - \pi}{1 - \pi}$, which is a simply a transformation of the model parameters, in a non-mixture cure model.

5.4.2 Flexible parametric cure models

In study IV of this thesis we use flexible parametric cure models to estimate statistical cure. We subsequently apply competing risks theory to get estimates of "personal cure". Flexible parametric cure models belong to the family of non-mixture cure models. They are also a special case of the flexible parametric survival model, outlined previously in equations (4-6). In contrast to other non-mixture cure models that have been adapted for relative survival, the use of flexible parametric cure models obviates the need to make strong distributional assumptions about the functional form of the survival function of the uncured patients and thus offer greater modelling flexibility [54]. It has also been shown empirically that the flexible parametric cure model tends to perform better than other commonly used cure models under scenarios where the survival of the patients is either relatively good (e.g., patients with localized melanoma) or poor (e.g., among elderly patients diagnosed with AML) [63]. These are situations where most cure models (adapted for relative survival) often fail to converge, or otherwise provide a poor fit to the data as compared graphically to empirical estimates of relative survival.

By definition, statistical cure occurs when the excess mortality, observed in the cancer patients, reaches zero. Such a scenario can be built into the flexible parametric survival model by constraining the log cumulative excess hazard function to have zero slope from a certain point in time after diagnosis. In practice, such constraint is imposed by treating the spline basis functions, $v_j(x)$, used to model the log cumulative excess hazard function in (4) in reverse order, and by restricting the parameter for the linear spline term to be zero. Thus, for $j = 1, \dots, K - 1$, $v_{j(x)}$ are now:

$$v_j(x) = \begin{cases} x, & \text{if } j = 1 \\ (k_{K-j+1} - x)_+^3 - \lambda_j(k_{max} - x)_+^3 - (1 - \lambda_j)(k_{min} - x)_+^3, & \text{if } j = 2, \dots, K - 1. \end{cases} \quad (20)$$

where $x = \ln(t)$ and $\lambda_j = \frac{k_{K-j+1} - k_{min}}{k_{max} - k_{min}}$. The parameter for the linear spline variable, γ_{01} , is then restricted to be 0.

The flexible parametric cure model provides an estimate of the relative survival, $R(t)$, via

$$R(t) = \pi^{F_Z(t)}, \quad (21)$$

where the cure proportion, $\pi = \exp(-\exp(\gamma_{00}))$, and the distribution function, $F_Z(t) = \exp(\gamma_{02}v_2(x) + \dots + \gamma_{0K-1}v_{K-1}(x))$.

When adding a set of covariates, \mathbf{z} , and D time dependent effects to the model, $R(t)$ becomes:

$$R(t; \mathbf{z}) = \exp(-\exp(\gamma_{00} + \boldsymbol{\beta}^T \mathbf{z}) \exp(\gamma_{02}v_2(x) + \dots + \gamma_{0K-1}v_{K-1}(x)) + \sum_{i=1}^D s(\ln(t); \boldsymbol{\gamma}_i)) \quad (22)$$

From (22) it follows that γ_{00} and $\boldsymbol{\beta}$ contribute to the modelling of the cure proportion, whereas parameters that depend on time are used to model $F_Z(t)$.

Flexible parametric cure models have been incorporated into the `stpm2` command in Stata [64]. This implementation was used in Study IV.

6 Study 1: Making cancer patient survival statistics more useful for patients and clinicians

6.1 Motivation

In the clinical research community there has been, and still is, much confusion about statistical methods for estimating survival in the presence of competing risks. As a result, a large number of papers have been published with the aim of clarifying when competing risks methods are needed, approaches to estimation, assumptions and interpretation of the estimated quantities, see for example [49, 65, 66, 67, 68, 69, 70, 71]. In this paper, a general aim was to clarify and discuss the relative merits of crude and net cancer patient survival in a population-based setting. In addition, we set out to demonstrate how period analysis, applied in a competing risk setting, can be used to predict 'absolute', (i.e. crude) survival probabilities applicable to newly diagnosed cancer patients. As a motivating clinical example, we assessed the impact of prognostic factors on the risk of prostate cancer death in relation to death from other causes than prostate cancer, and event-free survival, among recently diagnosed patients.

6.2 Data

The target group for the clinical application were all (recently diagnosed) men with intermediate or high risk localised prostate cancer in Sweden who are candidates for conservative or hormonal management. The study population was identified from the cohort of all ($n=99,051$) men with a recorded diagnosis of prostate cancer in the Swedish NPCR between 1996-2008. Included in the study were men with complete stage and treatment information ($n = 93,685$). Risk groups were formed according to a modified version of the National Comprehensive Cancer Network[®] classification, by combining the recorded TNM stage, tumour grade (according to the Gleason grading system), and serum Prostate Specific Antigen (PSA) level at the time of diagnosis. The intermediate risk category was defined as clinical local stage T1-2, N0, NX-, M0, MX and serum levels of PSA between 10-20 ng/ml or Gleason score 7, and the locally advanced high risk category as T3-4, N0, NX, M0, MX and/or PSA between 20-50 ng/ml and/or Gleason score 8 or higher. Patients not included in these risk group classifications, and patients who received curatively intended treatment were excluded, leaving 29,647 patients in the study population that formed the basis for the synthetic cohort for the period analysis. Figure 7 illustrates the entry criteria to the synthetic cohort. Under the chosen period window (2005-2009) no patients who died before 1 Jan 2005, ($n = 6,294$) contributed to the study-base. Thus, from the study population, 23,353 men were included in the analysis.

6.3 Methods

Our interest in this study was to estimate crude survival from prostate cancer (target event), versus crude survival from any other cause, by prognostic factors. A graphical representation of the com-

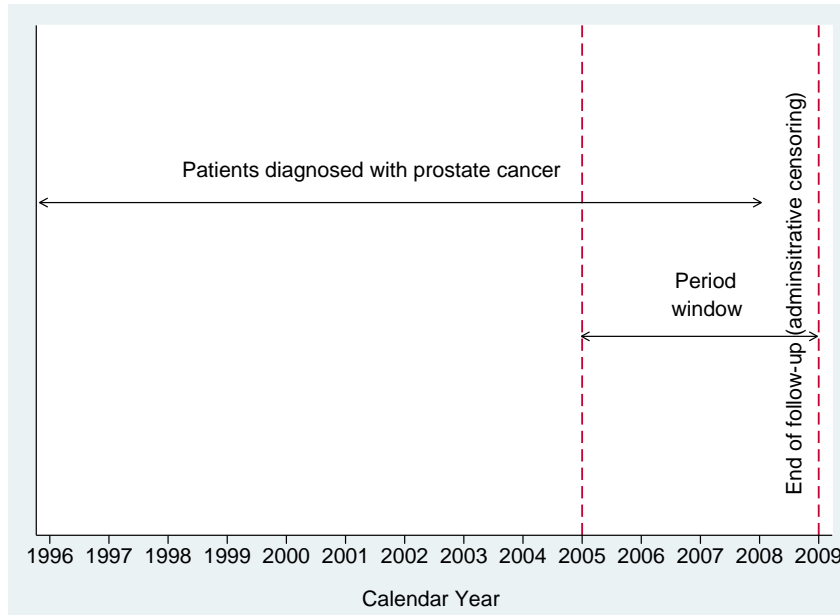


Figure 7 – Entry criteria to the synthetic cohort used in Study 1

peting risk model used in Study I is provided in Figure 8. We estimated the event-specific hazard function that corresponds to the excess hazard due to prostate cancer, $\lambda(t)$ using a flexible parametric survival model (with delayed entry), and obtained the expected hazard function in the absence of prostate cancer, $h^*(t)$ from population life tables. Crude probabilities of death were calculated by

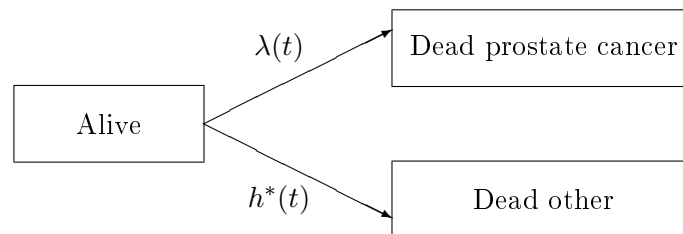


Figure 8 – Competing risk model used in study I

evaluating the integrals in equations (12) and (13) for $t \in [0, 10]$. The integrands were obtained using the predicted excess hazard, $\hat{\lambda}(t)$, and relative survival, $\hat{R}(t)$ from the flexible parametric model and by assuming that $S^*(t)$ and $h^*(t)$ were known (given sex, age and calendar year). To make the estimates of relative survival and excess mortality applicable to recently diagnosed patients we assumed that a period approach to modelling would give timely estimates of relative survival and excess mortality, and that the expected population mortality rates would remain constant 10-years into the future. At the time the study was carried out, the latest observed mortality rates available from the HMD were from 2009.

6.3.1 Approach to modelling

Table 1 contains a summary of all flexible parametric survival models that were fitted and contrasted in this study. In all models, 5 df were used to model the baseline excess hazard function. Likelihood ratio tests (LR-test) were used to formally test the significance of additional parameters (e.g., for time-dependent effects or interactions) to a given model. When modelling interaction effects between a covariate modelled using a restricted cubic spline (RCS) and a binary covariate, only the first spline term (i.e., the linear component) was included in the interaction. Model 8 was the final model and used to generate all results in Study I.

Model	Covariates included	Functional form	Df time-dependent effect	P-value from LR-test
1	Age at diagnosis	RCS 5 df	N/A	-
2	Age at diagnosis	RCS 5 df	3	< 0.001
3	Same as model 2 + Treatment	binary	N/A	< 0.001
4	Same as model 3 + Risk category	binary	N/A	< 0.001
5	Same as model 4 + Risk category	binary	3	0.0128
6	Same as model 5 + Treatment	binary	3	< 0.001
7a	Same as model 6 + Age x Risk category	RCS 1 df	N/A	0.5037
7b	Same as model 6 + Age x Treatment	RCS 1 df	N/A	0.0025
8	Age at diagnosis + Treatment + Risk category + Age x Treatment + Risk category x Treatment	RCS 5 df binary binary RCS 1 df RCS 1 df	3 3 3 N/A N/A	0.031

Table 1 – Summary of approach to modelling excess mortality

6.4 Findings and discussion

The idea behind this paper was to demonstrate how the combined use of competing risks theory, and statistical methods suitable to predict survival of newly diagnosed cancer patients, can provide data that can form a basis for risk counselling. Although such statistics are sought after by some patients and physicians, a responsible use of the information makes great demands on both its producers and consumers. At best, a carefully conducted statistical analysis of population-based data can provide a reasonable guess of survival estimates of newly diagnosed patients. At worst, it can mislead and give false hope to patients and their families. In this study, we provided a general discussion of the interpretation and assumptions of commonly reported statistics of population-based cancer patient survival. An overall aim was to raise the level of understanding for such statistics, and the discussion was centered around the following issues:

- Net and crude survival are conceptually different constructs that both have a place in cancer

care programmes, although the latter is more useful in the communication of risk between clinician and patient.

- A prediction model is only as good as the information that goes into it. In principle, individual-level prediction is the goal. In practice, the data available for modelling in population-based health registers are limited, and results must typically be interpreted on a group level.
- The manner in which survival statistics are presented to patients and their families, and the tools that are used for the purpose of conveying complex information, are instrumental in risk counselling.
- Careful consideration of the limitations that arise from statistical assumptions must be made when interpreting the results of any statistical analysis. In the setting of this study, these include assumptions made in relative survival versus cause-specific survival, sensitivity to modelling assumptions (inclusion of relevant prognostic factors, consideration of interaction effects, functional form of covariates, departures from the proportional hazard assumption, and the chosen approach to project the prognosis of patients).

The above points were highlighted in the context of prostate cancer patient survival in Sweden. In recent years, scientific papers that report estimates of crude prostate cancer survival have, in fact, become quite common [72, 73, 74, 75, 76]. We believe that the results from such reports are considerably more useful in a clinical setting than estimates of net survival, which do not provide an ideal basis for risk counselling as they are only interpretable in a hypothetical world (in which the patients do not live).

One aspect of risk counselling involves providing tools that can help patients make an informed decision about their preferred medical treatment. The use of observational cancer data for this aim is, nevertheless, suboptimal and results of the type presented in this paper should not be used in isolation, but in combination with evidence from randomized clinical treatment trials. For illustration, in the past 15-20 years there has been a rapid increase in (primarily) low, intermediate, and high risk localized prostate cancer, where a large contribution follow from the increased use of the PSA test [77, 78, 79]. Because the uptake of the PSA test, and the diagnostic work-up is skewed towards men who are healthier than the general population, strong selection mechanisms gave rise to our cohort [72]. This has consequences for the validity of relative survival analyses which assume that the patients are exchangeable to the general population (free from prostate cancer). As a means to circumvent this problem, our study did not include men with low-risk prostate cancer, a group where estimates of relative survival were > 1 for certain combinations of the covariates. However, it is possible that men who are at lower risk of dying of all-causes than the population at large are also, to some extent, overrepresented among men with intermediate and high risk localized prostate cancer. In addition, the results in this study indicate the presence of another selection mechanism that relates to how patients are assigned to treatment. When comparing, men in the same risk group and of the same age, men who receive hormonal therapy have a dismal prognosis compared

to men who are managed conservatively. This is probably best explained by the fact that men who are assigned to hormonal treatment at diagnosis are generally in poorer health compared to men who are eligible for watchful waiting (i.e., deferred hormonal treatment until clinical progression) and active surveillance (i.e., deferred curative treatment until perceived disease progression). The strong selection to the cohort and the indication to treatment warrant additional investigation and consideration of host-factors that predict all-cause survival. This can, for example, be done by expanding population life tables by indices such as socioeconomic position or level of comorbidities. Another way to achieve a more appropriate match of patients to the general population is to use the biological age (as opposed to chronological age) of patients as matching factor in the analysis. The estimation of biological age involves consideration of comorbidities in the group of patients, but obviates the need to ascertain the level of comorbid conditions in the entire back-ground population. If it is believed that the cause of death classification of prostate cancer patients is reliable, a third alternative is to estimate cause-specific survival, i.e., make explicit use of the recorded cause of death in the analysis (as opposed to estimate relative survival).

7 Study 2: Partitioning of excess mortality in population-based cancer patient survival studies

7.1 Motivation

An attractive feature of relative survival is that it provides an estimate of the excess mortality associated with a diagnosis of cancer. Excess mortality captures both the *direct* and the *indirect* contribution of cancer to mortality, as opposed to cause-specific survival which is limited to measure the direct contribution. Direct causes are all causes of death that would typically be classified as death from the cancer in question on the death certificate and include, for example, malfunctioning and failure of vital organs where the cancer was originally situated (or metastasized to) and cachexia. Indirect causes of death include late effects from treatment, e.g., cardiovascular complications, secondary malignancies, infections (that may be too hard for someone with terminal cancer to fight), and suicides. Late adverse health effects in cancer patients is a growing problem given the longer survival seen for most cancers. For example, in breast cancer survivors, several studies have identified an increased risk of cardiovascular disorders, mainly myocardial infarction, possibly associated with radio- and chemotherapy such as anthracyclines [80, 81]. These deaths are not straight forward to study since:

1. the physician that signs the death certificate can never know for certain if a cardiac event was induced by a treatment administered many years in the past or not.
2. indirect deaths are not immediately identifiable using a standard relative survival analysis (which provide an estimate of the overall excess mortality associated with cancer).

In this study we extended flexible parametric survival models for relative survival, with the aim to partition the overall excess mortality from cancer into two component parts; excess mortality from diseases of the circulatory system (DCS), and remaining excess cancer mortality.

7.2 Data

The method is illustrated using data on 70,655 women diagnosed with breast cancer in Sweden between 1973 and 1992 obtained from the Swedish Cancer Register. The restriction to diagnoses before 1993 ensured that all women had a potential follow-up of at least 15 years. Among these women, there were 40,361 death in all, of which 8,939 were classified as any disease of the circulatory system (according to the information stated on the death certificate).

The extended model requires population life tables that are stratified on broad categories of cause of death. These were constructed using individual-level data containing information about year of death, age at death, sex, underlying cause of death for all people who died in Sweden between 1961 and 2007. Two sets of life tables were constructed, one corresponding to the expected mortality rates from DCS, and one corresponding to the expected mortality from causes other than DCS. Data on population size, N , the total number of deaths, d , in Sweden and the total number of

DCS deaths, d_{DCS} , for the years 1973-2007 were obtained by year of occurrence from the HMD and the National Board of Health and Welfare, and were collapsed over sex, age and calendar year. Probabilities of death due to DCS and non-DCS were estimated by taking the ratio of the death counts and population at risk in matched intervals of age (i), sex (j), and calendar year (k). The corresponding mortality rates were then calculated using

$$h_{DCS,i,j,k}^* = -\ln\left(1 - \frac{d_{DCS,i,j,k}}{N_{i,j,k} - \frac{d_{DCS,i,j,k}}{2}}\right) \quad (23)$$

and

$$h_{other,i,j,k}^* = -\ln\left(1 - \frac{d - d_{DCS,i,j,k}}{N_{i,j,k} - \frac{(d - d_{DCS,i,j,k})}{2}}\right). \quad (24)$$

7.3 Methods

In a flexible parametric model (adapted for relative survival), the cumulative hazard at time t can be written as the sum of the cumulative expected hazard, $H^*(t)$, and the cumulative excess hazard, $\Lambda(t)$, according to expression (3). We now instead consider an alternative expression for $H(t)$,

$$H(t) = H_{DCS}^*(t) + H_{Other}^*(t) + \Lambda_{DCS}(t) + \Lambda_{Other}(t). \quad (25)$$

where $H^*(t)$ has been partitioned into the expected mortality from DCS, $H_{DCS}^*(t)$ and other causes than DCS, $H_{Other}^*(t)$, and $\Lambda(t)$ into the excess mortality from DCS, $\Lambda_{DCS}(t)$, and other causes than DCS, $\Lambda_{Other}(t)$, respectively. $H_{DCS}^*(t)$ and $H_{Other}^*(t)$ are assumed to be known from population life tables, and $\Lambda_{DCS}(t)$ and $\Lambda_{Other}(t)$ are estimated from a flexible parametric survival model. Thus, our interest is now in modelling the time to death from cancer induced DCS, and the time to death from other (non-DCS) cancer related causes. The corresponding excess hazard rates are depicted in Figure 9, together with a third outcome of potential interest, death from causes unrelated to the diagnosed cancer.

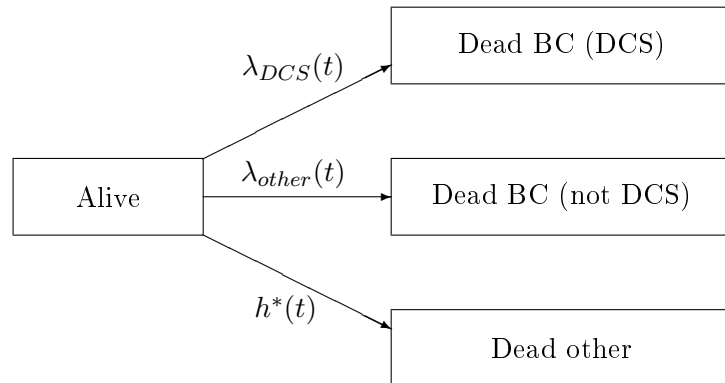


Figure 9 – Competing risk model used in study II

Important to remember is that, under this model, expected deaths from DCS are included in the

box that represents death from other causes than BC. The DCS deaths that occur as a consequence of treatment only constitute a small proportion of all DCS deaths combined.

In principle, $\lambda_{DCS}(t)$ and $\lambda_{other}(t)$ can be estimated using two separate flexible parametric models,

$$\ln(\Lambda_j(t; \mathbf{x})) = s(\ln(t); \gamma_{0,j}) + \mathbf{x}^T \boldsymbol{\beta}_j, \quad j \in \{cvd, other\} \quad (26)$$

where $s(\ln(t); \gamma_{0,j})$ provides an estimate of the log cumulative baseline excess mortality for cause j , and the vector $\boldsymbol{\beta}_j$ represents the covariate effects on cause j . However, an alternative approach to modelling, that allows for more flexibility in modelling the effects of covariates, is given by,

$$\ln(\Lambda_j(t; \mathbf{x})) = s(\ln(t); \gamma_0) + \mathbf{x}^T \boldsymbol{\beta} + c_j(\beta_{cvd} + s(\ln(t); \gamma_{cvd}) + \mathbf{x}^T \boldsymbol{\beta}_{cvd}) \quad (27)$$

where

$$c_j = \begin{cases} 0, & \text{if } j = \text{other}; \\ 1, & \text{if } j = \text{DCS}. \end{cases}$$

Using this approach, the two outcomes are estimated jointly by including a binary covariate, c , with cause of death information that enables estimation of separate baseline excess hazard functions for the different outcomes. The covariate effects, $\mathbf{x}\boldsymbol{\beta}_j^T$, are assumed to be common for the two causes of death, but this assumption can be relaxed by including additional interaction terms between a subset (or all) of the covariates and \mathbf{c} . The component-specific excess mortality rates, $\lambda_j(t)$, are related to $\Lambda_j(t)$ via $\lambda_j(t) = \frac{d}{dt}\Lambda_j(t)$.

Before a model of type (27) can be fitted, the data set must be transformed into *long format*. This is a common way to structure data sets prior to competing risks analyses, and involves assigning each individual as many observations in the new data set as there are outcomes under investigation. Moreover, a covariate that indicates cause of death is needed, as well as a death indicator for the specific cause. An example is shown below.

```

+-----+
| id   entry   exit  dead cause   rate  sex  ageddiag |
+-----+
| 5464   0       20   0   DCS   .00038  2   23 |
| 5464   0       20   0  Other  .00120  2   23 |
+-----+
| 6851   0   .74590164  0   DCS   .00109  2   57 |
| 6851   0   .74590164  1  Other  .00427  2   57 |
+-----+
| 7129   0  7.6939891  1   DCS   .00026  2   34 |
| 7129   0  7.6939891  0  Other  .00153  2   34 |
+-----+

```

In this sample data set, individual with $\text{id} = 5464$, was followed for 20 years and was still alive at the end of follow-up. Individual 6851, on the other hand, was followed for 0.75 years after which

she (sex = 2) died from a cause unrelated to DCS. Similarly, individual 7129 was followed for 7.69 years before she died from some disease of the circulatory system. In this data set, *rate* denotes the age-, sex- and calendar matched expected mortality rate from DCS and other causes, respectively. Models of type (27) estimate the component-specific³ excess hazard rates. Similar to a 'standard' competing risk model, as described in equation (12), the component-specific excess hazard rates can be used to calculate crude probabilities of death. The extension from (12) to component-specific crude probabilities of death is immediate

$$Cr_{can,j}(t) = \int_0^t S^*(u)R(u)\lambda_j(u)du, \quad j \in \{cvd, other\} \quad (28)$$

but requires the additional step of calculating $R(u) = \prod_j R_j(u)$.

7.4 Findings and discussion

In this study we present a number of summary measures for quantifying the risk for death from late effects of treatment relative to the overall risk of dying of breast cancer, or causes unrelated to the cancer. Component-specific excess mortality rates (and rate ratios) were used to assess the impact of prognostic factors on excess DCS mortality. Without making an assumption about independence (i.e., between deaths from breast cancer, both DCS and non-DCS, and other causes) these cannot be interpreted as net mortality rates from excess DCS. Instead, they describe the excess rate at which women with breast cancer die of DCS, in the real world where other causes of death also exist. Although estimates of net survival (or net mortality) are the gold standard for epidemiological studies into the causes and risk of cancer, in practise, the independence assumption is often not satisfied, and inference is then based on the real-world mortality rates. There is, nevertheless, no one-to-one relationship between the latter and crude survival probabilities. Competing risks methodology is therefore required to estimate the (real-world) survival probabilities from different causes of death. Figure 10 summarizes temporal trends in the 15-year crude probabilities of death from different causes among women diagnosed with breast cancer at ages 55, 65 and 75 respectively. Overall the probability of treatment-related DCS is very low (< 0.05) in relation to the probability of dying from other causes (from breast cancer, or unrelated to breast cancer).

We assume that the excess DCS mortality is associated with the cancer in question and, more specifically, that the excess DCS mortality is induced by the treatment. This is not a testable assumption, but its validity relies primarily on:

1. The quality of the information stated on the death certificates. Even though the excess DCS mortality rate is estimated in a relative survival framework, we must broadly categorise all underlying causes of deaths into either DCS or non-DCS in order to identify the deaths that occur *in excess* to what is expected in a comparable group of women free from breast cancer.

³These are sometimes referred to as cause-specific hazard rates [49] or transition intensities [82] in traditional competing risk and multistate model literature

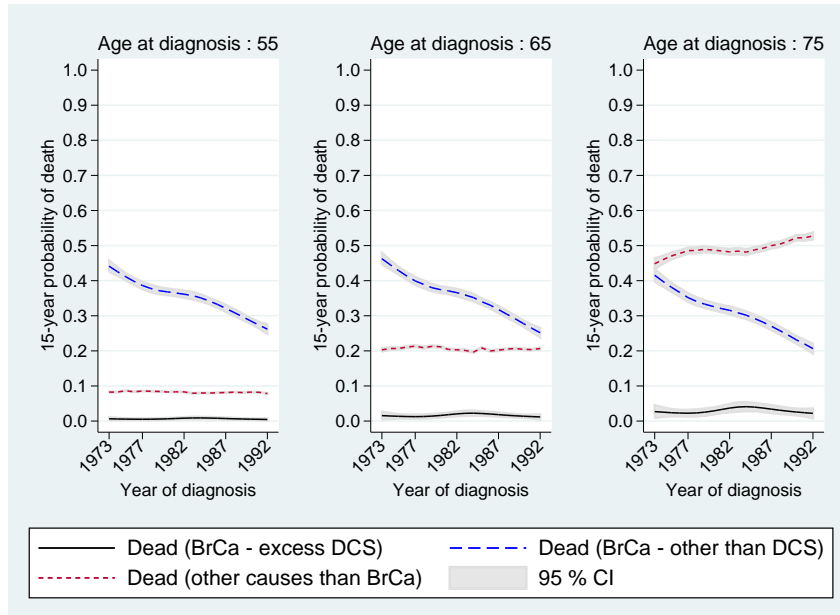


Figure 10 – Temporal trends in component-specific crude probabilities of death among women diagnosed with breast cancer in Sweden between the years 1973 and 1992

2. Whether women with breast cancer are exchangeable to the matched general population with respect to their cardio- and cerebrovascular risk profile, i.e., if the fact that the women have breast cancer is the only thing that distinguishes them from the general population with respect to their risk of developing, and dying from DCS.

Sensitivity analyses, explained in detail in the paper, were carried out and promoted to assess the impact of miss-classifications of the underlying causes of death on the estimated excess DCS mortality rates. In addition, we also estimated excess DCS mortality by laterality. This analysis was not included in the published paper, but we believe it provides support for the validity of our approach to estimating excess DCS mortality. It is known that the myocardial exposure of breast radiotherapy is higher among women treated for breast cancer in the left breast, as compared to the right breast [83, 84, 85]. Figure 11, shows that this pattern is also present in our data. As expected, the excess DCS mortality among women treated for breast cancer in their left breast is higher, and increases with elapsed time since diagnosis. Women treated with breast cancer in their right breast are also subject to excess DCS mortality (at a lower, and relatively constant, rate). We speculate that this is indicative of two driving mechanisms for the excess DCS mortality, one that arise from exposure to radiotherapy (left breast), and another that arise from chemotherapy related cardiac dysfunction, which become clinically apparent earlier than radiation damages [80]. The excess DCS mortality that is observed in direct relation to the diagnosis is, however, unlikely to be attributable to treatment, but possibly a severe side-effect from the high levels of psychosocial stress that is induced by the cancer diagnosis itself [86].

From a public health perspective, being able to study if changes in clinical practise towards reducing treatment-related mortality have had an impact on patient survival is important. Cardio-

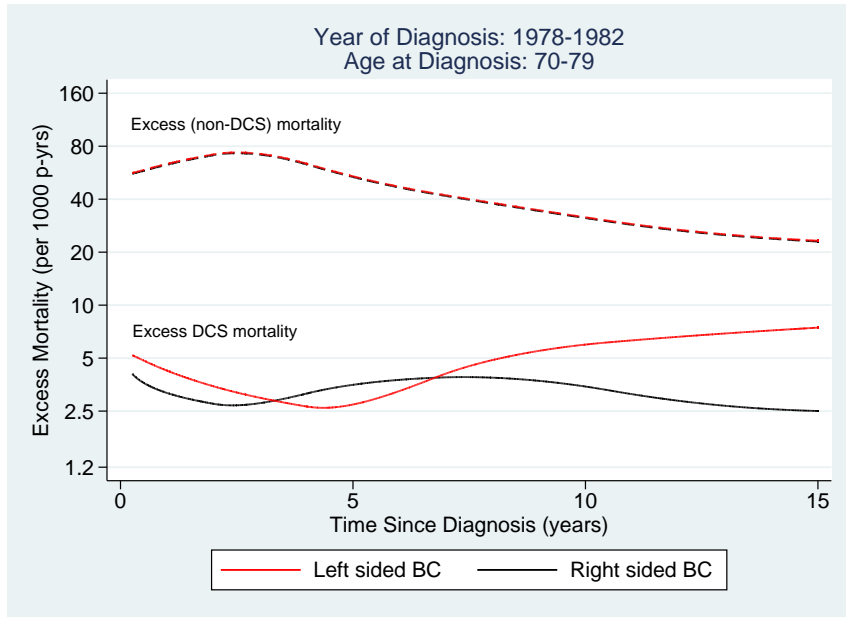


Figure 11 – Timing of excess non-DCS and excess DCS mortality by breast cancer laterality among women diagnosed in Sweden between the years 1978 and 1982 at ages 70-79.

and cerebrovascular late effects from treatment is merely one indirect cancer consequent death out of many possible, for example, risk for death from secondary cancers and infections in patients with Hodgkin lymphoma, or suicide in patients with e.g., lung, prostate, or head and neck cancer. Finer partitioning of excess cancer mortality would be useful to investigate the relative contribution of each such cause to the overall risk of dying from the diagnosed cancer. Such analysis would require additional use of the information stated on the death certificates and further work should address under what situations finer stratification is feasible.

8 Study 3: Temporal trends in mortality from diseases of the circulatory system after treatment for Hodgkin lymphoma

8.1 Motivation

Survival after Hodgkin lymphoma has increased substantially in the past four decades and, for patients aged less than 65 years at diagnosis, the disease is now highly curable [87]. The improved prognosis is likely attributable to improved patient assessment and staging, the development of effective multi-agent chemotherapy, introduction of combined-modality therapy with reductions in radiation field size and dose, and more apt evaluation of treatment response. The aim of this study was to apply the methodology developed in Study II to present clinically interpretable estimates of temporal trends in the burden of excess mortality from diseases of the circulatory system (DCS) among Hodgkin lymphoma survivors who were treated in the 1970's through 1990's. In addition, since the treatments administered to these patients are, to some extent, considered outdated in modern Hodgkin lymphoma management, we also aimed to predict the future clinical burden among patients diagnosed and treated in more recent years.

8.2 Data

Included in the study cohort were all patients diagnosed with Hodgkin lymphoma in Sweden between 1973 and 2006 at ages 19 to 79 ($n=5,832$). These patients were identified in the Swedish Cancer Register using ICD7 code 201. No patients were excluded on the basis of previous cancer diagnoses, although in patients with multiple primary Hodgkin lymphoma diagnoses, we only considered the first diagnosis ($n=16$ excluded). Patients reported as having suspected, or not microscopically verified Hodgkin lymphoma were excluded ($n = 36$). Likewise were patients with a Hodgkin lymphoma that was diagnosed incidentally at autopsy ($n=318$), leaving 5,462 patients in the analysed cohort. The prognostic factors under investigation were sex, age at diagnosis and calendar year of diagnosis.

8.3 Methods

8.3.1 Approach to modelling

We estimated the component-specific excess DCS mortality rate, and remaining (non-DCS) excess mortality using flexible parametric survival models of the type described in Study 2. Each component-specific baseline excess hazard function was modelled using 5 degrees of freedom (df). The effects of age and year were modeled continuously and non-linearly using restricted cubic splines, where the knots were placed at the 5th, 25th, 50th, 75th and 95th percentiles of the distribution of the two variables, respectively. Moreover, the effects of these covariates were estimated independently between the outcomes (i.e., no main effects were assumed to be shared between the two outcomes) and time-dependently. Time-dependent effects were modeled with 12 df (3 df x 4 parameters representing the basis vectors for the age and calendar effects respectively). The assumption

that the time-dependent effect of calendar year was shared between the outcomes was relaxed by including a three-way interaction term (between calendar year, cause of death and the time scale). In all, 57 parameters were estimated in the final flexible parametric model. When investigating the statistical significance of interaction terms, we sometimes kept non-significant terms in the model in order to avoid imposing constraints on estimates that were of primary interest. For example, we did not constrain the effect of calendar year to be the same for both the excess DCS mortality and the remaining excess mortality since temporal trends in these rates were the main interest of the study.

8.3.2 Predicting excess mortality and crude probabilities of death into the future

We were only able to fully observe the 20-year excess DCS mortality rate, and remaining excess mortality rate for patients diagnosed prior to 1988 (due to administrative censoring at 31 Dec 2007). For patients diagnosed from 1988 and onwards, follow-up was, nevertheless, partially observable, and used to extrapolate the long-term excess mortality beyond the range of existing data. The extrapolations were based on the fitted trend (within the observable data) from the model specified above. Figure 12 shows how the model-based extrapolation of the 20-year excess DCS mortality behaves beyond the observable data for patients diagnosed at age 60 and in different years. Apart from any assumptions that were made in terms of specifying a model for the covariate effects, the extrapolated rates should only be interpreted under the additional assumption that no unpredictable changes in the general pattern and trends of the mortality of either the HL patients or the general population occurs.

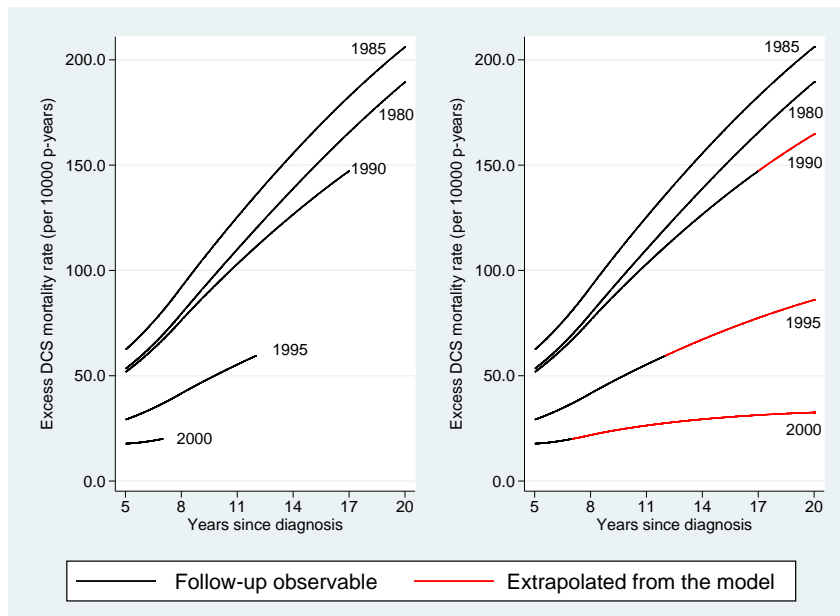


Figure 12 – Illustration of the estimated excess DCS mortality predicted within the range of observable follow-up and extrapolated outside the range of follow-up for male patients, aged 60 years at diagnosis, and diagnosed with Hodgkin lymphoma in Sweden between 1980 and 2000.

8.4 Findings and discussion

Although excess mortality from DCS is a concern for long-term Hodgkin lymphoma survivors, the actual risk of dying from such treatment related side-effects is small in relation to the risk of dying from the underlying disease, or from other causes (unrelated to the disease). The 20-year excess DCS mortality has decreased continually since the mid 1980's, and is predicted to further decrease among patients diagnosed in the modern era. There are many possible explanations for these findings. The National Care Programme has regularly been updated with regards to principles for treatment, through restricting to low-dose radiotherapy and reducing and alternating chemotherapy cycles, with the aim of minimizing the risk for long-term adverse events, without compromising the anti-tumour effects [88]. However, lack of treatment data in this study, prevented us from disentangling the relative contribution of these efforts. Other actions to reduce treatment-related cardiotoxicity have also included patient-specific recommendations and surveillance visits to estimate the individual risk profile and enable early detection of symptoms [89].

Few studies on this topic have presented crude probabilities of death as measure of the risk that Hodgkin lymphoma patients face to die of late cardiotoxic effects from treatment. Moreover, many authors have recognized the difficulty in estimating the effects of long-term treatment complications in patients diagnosed and treated more recently [90, 91, 92]. This study is novel in the sense that we have strived to do both these things. Although the panorama of possible late effects attributable to modern Hodgkin lymphoma treatment are yet to be verified once the data becomes available, we argue that, meanwhile, this study maximizes the amount of information that can be obtained from the available data.

The predicted 20-year crude probabilities of death for patients diagnosed between the years 1988 and 2003 were calculated under the assumption that there will be no changes in the all-cause population mortality between the years 2007 and 2023. However, the average life span in Sweden has increased by 2.5 years, on average, per decade since the 1900 [4]. Of all causes that contribute to mortality in the population, mortality from DCS has decreased most, for both men and women. The relative reduction in DCS mortality, for individuals aged 30 to 79, over the past three decades has been approximately 3-4% per year. In order to assess the sensitivity of the published 20-year crude probabilities of death for patients diagnosed from 1988 and onwards, we also re-did the analysis using a set of population mortality rates projected for the years 2013-2023 by Statistics Sweden, using the so called Lee-Carter model [93]. This model incorporates past mortality trends and has become the leading statistical model for population mortality forecasting, both in the demographic literature and in practical applications. The Lee-Carter projections are summarised in Figure 13, whereas Table 2 displays the flat mortality rates that were applied in our study. The projected mortality rates for individuals up to age 60 are close to those estimated from the Lee-carter model. However, for individuals older than 60, the rates used in our study are well above the rates projected by Statistics Sweden.

Sex	Age					
	30	40	50	60	70	80
Males	0.7	1.1	3.2	7.1	20.6	66.5
Females	0.3	0.7	2.0	4.8	12.2	44.6

Table 2 – Projected mortality rates (per 1000 person-years) used in study III (assumed fixed for the years 2008-2023)

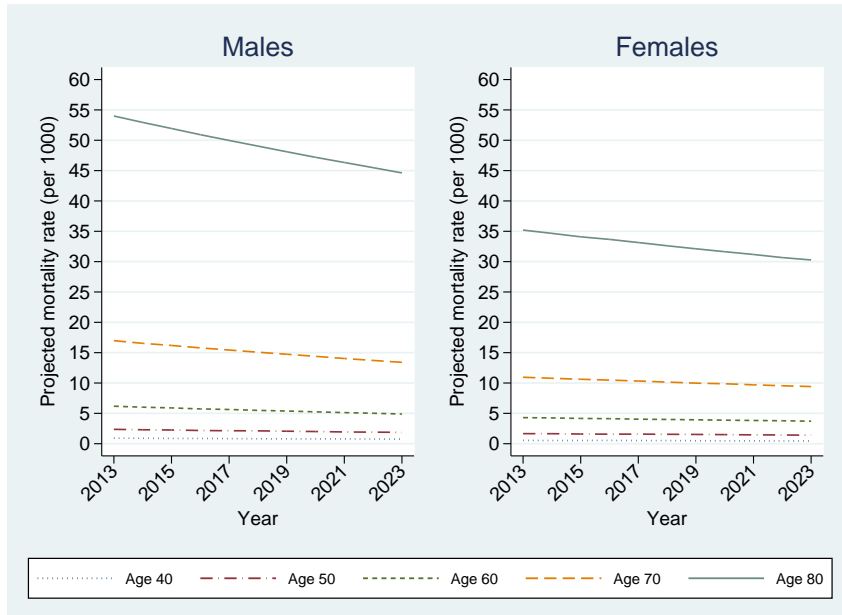


Figure 13 – Projected mortality rates (per 1000 person-years) for the Swedish population between the years 2013-2023 using the Lee-Carter model. Source: Statistics Sweden 2013-08-06

The crude probabilities of death from different causes that were published in study III are repeated in Figure 14 (left-hand side). In addition, Figure 14 shows the difference (expressed in percentage points) between the published results and the results obtained when population mortality rates from the Lee-Carter models have been applied instead. The published crude probabilities of death from other causes than Hodgkin lymphoma are overly pessimistic, as compared to those that were obtained using the mortality rates projected by Statistics Sweden using the Lee-Carter model. The difference between the two approaches was greatest for patients who were 60 years or older at diagnosis and diagnosed in the most recent years (overestimated by 5 percentage points for patients diagnosed in 2003). Whilst the direction of these findings on deaths from other causes than HL are consistent with what we would expect to observe if the all-cause mortality rates were to decrease in the future, changing to the Lee-Carter projected mortality rates appears to have a negligible impact on the crude probabilities of death from Hodgkin lymphoma (DCS or non-DCS).

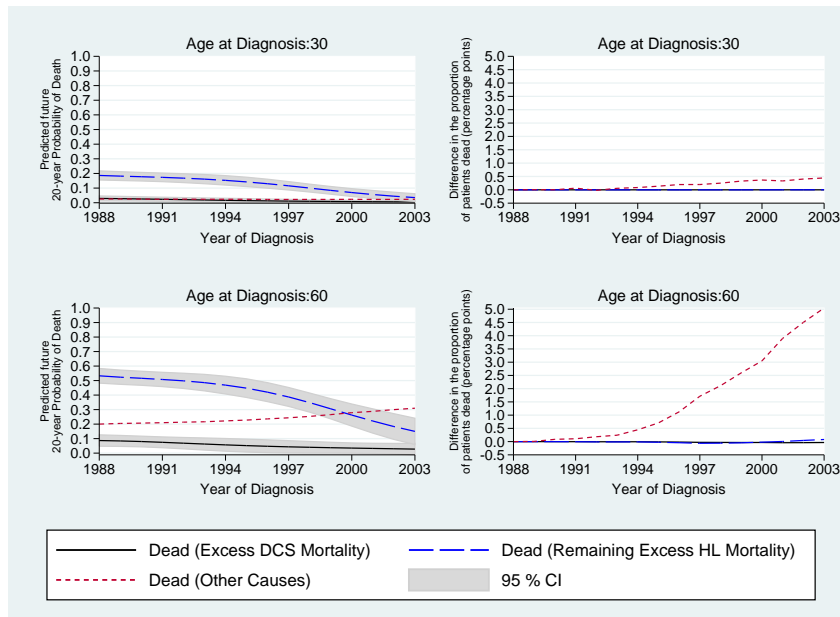


Figure 14 – Crude probabilities of death (as published), and the difference (expressed in terms of percentage points) between the published results and the results obtained when population mortality rates from the Lee-Carter models have been applied.

We, nevertheless, recommend further investigation into the topic and impact of different assumptions for projecting the expected mortality rate in future studies of this kind. A less naive approach than that applied in the current study might be warranted in other cohorts of patients, particularly if the interest is in elderly patients.

9 Study 4: The application of cure models in the presence of competing risks

9.1 Motivation

Estimates of cancer survival in the presence of competing risks are frequently used to relate the patients' risk of dying from their cancer to that of dying from competing causes, or to still be alive, as a function of elapsed time since diagnosis. A natural question for patients to ask when faced with such statistics is related to their chance of being cured from the diagnosed cancer.

In this study we demonstrate how flexible parametric statistical cure models, combined with competing risks theory, can be used to estimate so called 'personal cure' in population-based cancer patient survival studies.⁴ We also show how patients who are still alive, and under follow-up, can be partitioned into two groups as depicted in Figure 15.

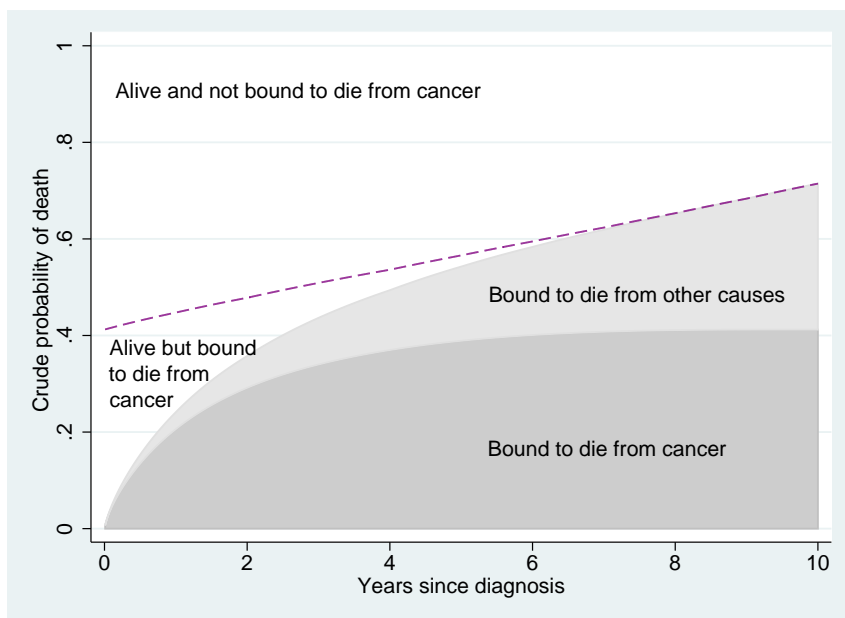


Figure 15 – Partitioning of the crude probability space of cancer patients

Furthermore, we derive and discuss the usefulness of conditional crude probabilities as a tool for supplying patients with updated estimates of their anticipated prognosis.

9.2 Data

All cases of malignant melanoma (ICD7: 190x), colon cancer (ICD7: 153x) and acute myeloid leukemia (AML)(ICD7: 205.0, 205.9, 206.0 and 206.9) diagnosed in Sweden between 1973 and 2007, at ages 19 to 80, were extracted from the Swedish Cancer Register. Information about date of death for deceased individuals was retrieved from the Swedish Causes of Death Register. No

⁴Even though the term 'personal cure' has been introduced in a population-based setting previously, using a grouped data approach [94], we have deliberately avoided to use this terminology in the current study since it provides a poor description of what is, in fact, estimated.

patients were excluded on the basis of previous cancer diagnoses, although in patients with multiple tumors of the same type, we only considered the first diagnosis. Incidental autopsy findings and benign cancers were excluded from each of the three cohorts (melanoma: 11,958 exclusions, colon cancer: 14,517 exclusions, AML: 2,436 exclusions). The last observable date for administrative follow-up was December 31, 2012.

9.3 Methods

The general approach to estimating the quantities of interest in this study is outlined below:

1. Estimate the (net) survival from flexible parametric cure models. This involves defining a point from which the patients are assumed to no longer experience excess mortality from the diagnosed cancer. In this study, and for all three cancer types, statistical cure was assumed to occur at 10 years after the cancer diagnosis.
2. Transform estimates of (net) survival from the flexible cure model into crude probabilities of death from cancer by evaluating:

$$Cr_c(t) = \int_0^t S^*(u)R(u)\lambda(u)du,$$

and

$$Cr_o(t) = \int_0^t S^*(u)R(u)h^*(u)du.$$

3. Estimate the asymptote, $P_{btd,c}$, to the cancer-specific cumulative incidence function, $Cr_c(t)$. Due to the pre-specified time of cure, this asymptote corresponds to $P_{btd,c} = Cr_c(10)$, and defines the proportion of patients who are bound to die from their diagnosed cancer, (i.e., one minus the 'personal cure' proportion).
4. Calculate cumulative crude probabilities that patients who have not yet died (from any cause) at time t after diagnosis will eventually die from cancer, $Cr_{alive,c}(t)$, versus other causes, $Cr_{alive,o}(t)$, using:

$$Cr_{alive,c}(t) = P_{btd,c} - Cr_{cr,c}(t) \tag{29}$$

and

$$Cr_{alive,o}(t) = 1 - Cr_{cr,o}(t) - P_{btd,c} \tag{30}$$

The above steps partition the crude probability space of cancer patients according to Figure 15. The proportion of patients who will not die from cancer, conditioning on having survived until time t , was calculated from:

$$Cr_{o|alive}(t) = 1 - \frac{P_{btd,c} - Cr_c(t)}{1 - Cr_c(t) - Cr_o(t)} \tag{31}$$

The `stpm2` program in the Stata software was used to fit the flexible parametric cure model. We wrote a post-estimation command that was subsequently used to estimate equations (29)-(31). The numerical approximations of the integrals and the variances used to construct 95% confidence intervals were performed using the methods described in chapter 5.3.1 of this thesis.

9.3.1 Approach to modelling

For the melanoma and colon cancer cohorts, we fitted flexible parametric cure models with 6 df for modelling the baseline cumulative excess hazard function (where the knots placed at the 0th, 20th, 40th, 60th, 80th, 95th and 100:th centiles of the distribution of the uncensored log survival times). For the AML cohort, we added one more internal knot at 8 years after diagnosis because the last internal knot would otherwise be placed too early during follow-up (at 4.1 years). It has been shown that it is important to disperse the internal knots over the entire follow-up when applying flexible parametric cure models in order to achieve a good fit to the data [64].

All models were adjusted for sex, year of diagnosis and age at diagnosis. The effects of the continuous variables were assumed non-linear and modeled using restricted cubic splines (using 4 df to represent the effect of age, and 3 df for the calendar effect). Furthermore, the effect of age and calendar year were modeled time dependently (using 3×4 df and 3×3 df, respectively, for the time-dependent effects), and assumed to interact (using an additional 3×1 df for the interaction effect).

9.4 Findings and discussion

To cure a cancer patient from her disease can be viewed as the ultimate goal in medical decision making, and the ability to quantify the anticipated survival from the prospect of cure should have high clinical value. Statistical cure models aim to estimate the proportion of cancer patients who will not experience any excess mortality from their disease. However, the cure proportion is estimated in the absence of competing causes of death, and as such, it does not represent the actual survival patterns in a given cohort of cancer patients. In this study, we have shown how to recalculate net survival, and the statistical cure proportion, from a flexible parametric cure model into crude survival probabilities, and a competing risks analogue to the cure proportion. In addition, we have shown how the crude probability of still being alive after a diagnosis of cancer can be further partitioned to estimate the proportion of patients who were alive and bound to die from their cancer, versus alive and bound to die from other causes. Results from this analysis, presented for males diagnosed with melanoma, colon cancer and AML, respectively, in Sweden in 2005, are summarized in Figure 16. These are cancer types where statistical cure models have previously been applied to Swedish data, but where competing risks were not taken into account in the analyses [95, 96, 97]. The interpretation of the results in the earlier studies was therefore somewhat different. Similar to other measures of net survival, the statistical cure proportion is an appropriate measure to evaluate the progress of control activities, or improvements of cancer treatment. As such, it is also useful for cancer patients, but in an indirect way. Crude estimates of survival reflect the survival patterns

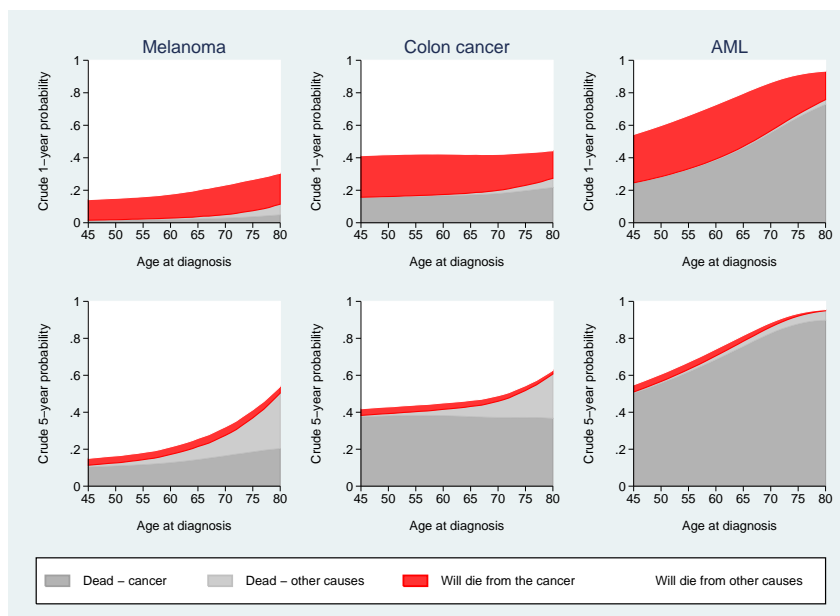


Figure 16 – Crude probability of death from melanoma, colon cancer, and AML in Sweden among males diagnosed in 2005.

that are actually observed. Given the increasing number of long-term cancer survivors, there is also a need for statistical survival measures that are applicable under the realistic situation where other causes than cancer may lead to the death of the patient. We believe that such measures are of a more direct interest to cancer patients and physicians for the purposes of risk communication, but also for scheduling follow-up visits to the clinic for long-term survivors.

Survival time from diagnosis is of great relevance to clinicians, researchers, and administrators, but is less relevant to patients who have already survived some initial period. Mortality for many types of cancer is highest in the months immediately following diagnosis, and consequently, survival estimates from diagnosis will not be of great relevance to patients who have already survived a number of months (or years). Of greater interest are so-called conditional survival probabilities; the estimated future (crude) survival probability, conditional on having survived some initial period following diagnosis. Conditional survival probabilities can be used to provide regular updates of the prognosis and can provide critical information that may decrease anxiety associated with the diagnosis of cancer, and help deal with feelings of uncertainty about the future. Baade and colleagues, for example, state that:

"for patients who survive 1,3 and 5 years after diagnosis, access to updated information for each point of their cancer recovery is important, providing scope for evidence-based optimism as they progress on the survivor's journey" [98].

Conditional crude survival probabilities provide such information but have not been reported often in practice. Figure 17 shows the age-specific conditional crude probabilities that patients who are still alive one year and three years, respectively, after their diagnosis, will die from a cause unrelated to their diagnosed cancer. For all three cancer types, a general improvement in prognosis follows

from each additional year survived. For colon cancer and melanoma, nearly all patients who have survived at least five years, irrespective of age at diagnosis, are predicted to die from some cause unrelated to their cancer.

A potential criticism of the results presented in Figure 17 is related to how sensitive the results are to the assumption that statistical cure has been reached after 10 years of follow-up. We strongly suggest that flexible cure models should only be used to model cancer survival in situations where it is known that excess mortality (on a group level) does eventually reach zero. For all cancer types included in this study, there have been previous reports about statistical cure. Even so, it is useful to also consider the robustness of the results with respect to the assumption that is made about the timing of statistical cure by bringing the point of cure further forward in time. In Figure 17 we have therefore also overlaid the results from such sensitivity analyses, where statistical cure was instead assumed to have occurred after 11, 12 and 13 years after diagnosis. The results are very close to those based on the assumption that cure was reached after 10 years.

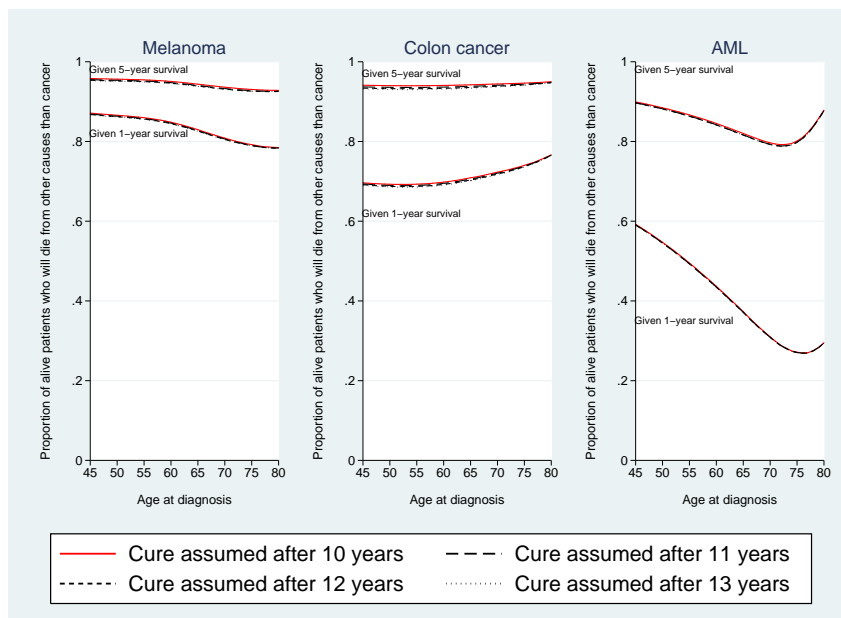


Figure 17 – Estimated proportion of patients with melanoma, colon cancer and AML who are predicted to die from causes other than their cancer, conditional upon having survived for one and five years, respectively.

10 Conclusions and discussion

The overall aim of this thesis has been to develop and apply statistical methods for presenting population-based cancer survival statistics in a manner relevant for physicians and patients. In one way or other, most statistical analyses of cancer patient data can probably be classed as relevant for physicians and patients. For example, randomized clinical treatment trials are necessary to evaluate the efficacy of new treatments and, if the outcome is satisfactory, the new treatment will benefit future patients, so the study is relevant for patients in that sense. An observational study that correlates survival of cancer patients with dietary habits can be relevant for patients if the knowledge gained provides clues into the aetiology of the disease, and can be used as grounds for nutritional recommendations that improve the well-being of the patients. Studies that evaluate the effectiveness of screening programmes are highly relevant for patients, since acceptance for, and participation in, existing programmes can be crucial to get an early diagnosis and to reduce the burden of cancer. All of these research questions would typically be investigated using survival analysis in a net framework⁵. However, throughout this thesis, I have referred to methods that are relevant for physicians and patients almost synonymously with methods that aim to predict the prognosis after a diagnosis of cancer in the presence of competing risks. Such methods are suitable to e.g., predict survival for newly diagnosed patients, and to balance the risk of death from cancer, overall or from treatment-related complications, against the risk of death from causes unrelated to the diagnosed cancer. By presenting results in a manner that is understandable for both health care professionals and lay audiences, and by explaining under what assumptions the methods and statistical analyses are valid, we believe that such methods can be of direct use in risk counselling between patients and physicians.

All studies in this thesis include competing risks methodology. In my experience, a wider recognition of situations when the statistical analysis needs to be made in the presence, as opposed to absence, of competing risks is needed. To date, many tutorial papers have been written on this subject, but the uptake of new methods in applied research is generally slow and there still seems to be much confusion surrounding this topic amongst medical researchers. In part, this is probably due to a more general lack of understanding of the theoretical quantities that are, in fact, estimated using traditional methods for survival analysis.

10.1 What do we aim to estimate?

To help researchers, unacquainted with statistical methods for competing risks, decide what analysis to do, and what quantities can be estimated, I have developed a decision-tree (see next page) that can serve as an informal guide towards an appropriate analysis. In principle, the choice of survival analysis depends on the target quantity of interest, in this example net or crude survival.

⁵By a net survival framework I include methods that aim to estimate (net) survival functions, hazard rates and/or hazard ratios

What is the aim of the study?

Aim: Hypothetical construct

Investigate a hypothesis that is best answered by quantifying the probability (or rate) of a specific event in the hypothetical scenario where the competing risks are removed.

Examples: Studies that pertain to questions about disease etiology, temporal trends or otherwise require comparisons of groups of patients (e.g. by hospitals, regions, countries etc.) where mortality from other causes may differ across time or groups.

Aim: Real world construct

Investigate a hypothesis that is best answered by quantifying the probability of a specific event in the real world in the presence of competing risks.

Examples: Studies intended for risk communication between patients and clinicians where the goal is to quantify the probability of experiencing the event of interest whilst adjusting for the competing risk mortality. Studies for e.g., resource allocation that require estimates of disease-specific survival in a well-defined cohort that takes into account that not all patients will live long enough to experience the event of primary interest.

Is the mortality from the competing cause (s) independent from the mortality from the cause of interest?
Note: This is not testable

Y

N

Net (marginal) survival and hazards can be estimated and the research question can be answered.

Is it possible to identify and adjust for the factor that induces the underlying mechanism for the dependence?

Y

N

Adjust for the relevant factor in order to achieve conditional independence between the competing events and the outcome.

Net (marginal) survival and hazards can be estimated and the research question can be answered.

Path 1

Cause-specific hazards and hazard ratios can be estimated.
These can often be used to answer the research questions.
Note: Kaplan-Meier estimates are biased and should not be interpreted.

Net (marginal) survival and hazards can be estimated and the research question can be answered.

Path 2

Special competing risk methodology required.
Two approaches available

Cause-specific hazards can be estimated using a standard approach to estimation.

Subdistribution hazards and ratios (SHR) can be estimated e.g., via Fine & Gray models.

No one-to-one relationship between the cause-specific hazards and cause-specific cumulative incidence function.
Must evaluate:
$$\int_0^t S(u)h_c(u)du$$

to get the cumulative incidence function (CIF).
The research question can be answered and differences between groups can be tested by calculating differences in CIFs.

Path 3a

One-to-one transformation of sub-distribution rates to the cause-specific cumulative incidence function.
The research question can be answered and differences between groups can be tested based on the SHR

Path 3b

Whether the fitted model actually provides a reasonable estimate of the target quantity depends on the degree of systematic and random error that is inherent in the data, or arise in the estimation. The flowchart aims to explain under what conditions net (marginal), cause-specific and subdistribution hazard rates can be estimated, and what type of research questions each of these constructs is best suited to answer⁶.

Broadly speaking, net survival is the underlying quantity of interest in many epidemiological applications (e.g., in studies of disease aetiology, or studies that investigate the effect of prognostic factors on cause-specific survival). Under the assumption of conditional independence between the competing events there is a one-to-one relationship between net hazard rates and net survival (path 1 in flowchart). This assumption is not formally testable using the competing risk data, but if violated, cumulative survival curves (i.e., Kaplan-Meier curves or relative survival curves) do not provide unbiased estimates of net survival, and the estimated hazard rates lose their (net) interpretation. Without the independence assumption, the hazard rates become cause-specific hazard rates which quantify the event rate in the realistic situation where individuals might experience a competing event before the target event (path 2 in flowchart). That is, net hazard rates and cause-specific hazard rates are technically estimated using the same statistical machinery, but the interpretation of the estimated hazard, as one or the other, depends on whether the independence assumption can be regarded satisfied or not. The cause-specific hazard rates (and rate ratios), nevertheless, still carry information that can be used for statistical inference. In practice, most epidemiological studies that aim to investigate a research question that pertains to net survival are probably analysed using cause-specific hazard rates, and not marginal hazard rates, since the former can always be estimated.

Crude survival, on the other hand, is primarily used to assess the impact of some risk/prognostic factor on a specific target event, in the real-world situation when competing events might preclude the event of interest from occurring. Unfortunately, the cause-specific hazard rate for a single cause has no exact mathematical relationship with the corresponding cause-specific crude cumulative survival (i.e., cause-specific cumulative incidence functions). Instead, there are two main approaches to estimation of crude survival from a regression modelling perspective,

1. by estimating the cause-specific hazard (or the excess hazard) and the all-cause survival and transforming them to the crude survival scale in a separate step using equation 8 (path 3a in flowchart).
2. by estimating the *subdistribution hazard*, e.g., using the Fine-Gray model, which has a one-to-one correspondence to the cause-specific cumulative incidence function (path 3b in flowchart).

In contrast to the estimation of net survival, the independence assumption is not required to achieve a valid estimate of crude survival. Throughout this thesis, we have applied the first approach to estimation. The Fine-Gray model has not been used, or discussed, in this thesis previously but is a

⁶Systematic errors, such as confounding, selection-bias, recall bias etc, are not included in the flowchart but must of course also be considered in order for any study to be valid.

commonly adopted approach to estimating crude survival. In fact, the Fine-Gray model is probably the most commonly applied method for estimating crude survival. The main reasons why we have not considered the Fine-Gray model in this thesis, despite it's wide use in applied research, include:

- We measure cancer-specific survival using relative survival, as opposed to cause-specific survival. The Fine-Gray model has not been adapted for relative survival.
- In addition to crude survival probabilities, we also studied excess hazard rate ratios as an effect measure for the association between certain covariates and net survival. For example, in studies II and III we compared and contrasted the influence of age at diagnosis, sex, and year of diagnosis, on the excess DCS mortality and remaining excess mortality respectively. These were estimated from the same flexible parametric model that gave the estimates of $S(t)$ and $\lambda(t)$ that were subsequently used to calculate crude survival probabilities. Subdistribution hazard rate ratios, estimated from a Fine-Gray model have a different interpretation, as compared to excess hazard rate ratios, and cannot be used to answer research questions that pertain to net survival.
- Our group has developed a range of statistical methods for population-based cancer patient survival by extending flexible parametric survival models. A nice feature of the methods described in studies II and V, is that they are also fitted within this modelling framework. We hope that this fact increases the acceptability and the uptake of the methods in applied research.

Study I can be viewed as a classical competing risks analysis (in a relative survival framework) with the goal to quantify probabilities of death from prostate cancer whilst adjusting for the competing risk mortality. Studies II-IV, on the other hand, extend existing competing risks methodology, by adapting and applying it in a new setting, in order to provide additional insight into the anticipated prognosis of cancer patients.

For example, partitioning of excess mortality into component parts is conceptually similar to a cause-specific competing risks analysis, where treatment related DCS deaths and other (non-treatment related DCS) deaths are treated as separate outcomes. However, in practice, the fact that treatment related DCS deaths are not possible to distinguish from DCS deaths that would have occurred irrespective of whether the patient was treated or not, makes the analyses in studies II and II impossible to carry out in a cause-specific setting. Thus, carrying over, and translating well-known methods from the competing risks literature from a cause-specific to a relative survival framework was needed.

Similarly, in study IV, we adapted competing risk methodology for relative survival to incorporate the assumption that a proportion of the patients will eventually be statistically cured from their disease. Such adaptation enabled estimation of "personal cure", and further partitioning of alive cancer patients into distinct groups, based on their risk of eventually dying from the diagnosed cancer, or from some cause unrelated to their disease. Summary measures of cancer patient survival

that incorporate the prospect of cure are rarely reported, but can be useful in the dialogue between physician and patient since it puts the health risks, faced by the patients with a curable cancer, in context.

10.2 Presentation of results

With the development of new statistical methods, aimed at producing survival statistics that describe the likely outcome of a given disease, come challenges related to the presentation of the results from such analyses.

In epidemiological association studies, there are often one or two variables (and possibly the interaction between the two) whose effect on the outcome is of main interest. However, additional covariates are typically also included in the analyses to control for potential confounding. The results of such studies are often quite easily summarized in terms of hazard ratios in tabular format. In studies aimed at prediction, more complex/larger statistical models are typically required in order to give meaningful predictions. In principle, any pertinent data useful for predicting the outcome should be considered in predictive models, although attention must also be paid to avoid overfitting. In practice, at least when it comes to register-based studies, we use the data that we have at hand for prediction. In the majority of the studies in this thesis, the data available were those recorded in the Swedish Cancer Register. This limited the covariates available for analysis to relatively basic demographic data about the patient (age at diagnosis, sex, year of diagnosis etc.), and hardly any data related to the tumour or treatment (except in study I where data from the NPCR were used). Even so, when continuous covariates were modelled continuously, as opposed to categorically (which seems to be the tradition in epidemiology) the number of possible covariate patterns grew quickly, even with as few as 3-4 predictors in the model. A structured presentation of results becomes still harder when considering conditional estimates of survival, since updated predictions should probably be calculated for several time points after diagnosis in order to have value in a clinical setting. In order to comply with the space limitations set up by the scientific journals in which our studies were published, we focussed on demonstrating the statistical methods and thus restricted the presentation of results to "typical" patients. New modes/formats for displaying predictions are needed for a more exhaustive presentation of results from studies of this type, a fact that has also been addressed by researchers at the Norwegian Cancer Registry as follows:

Providing realistic, relevant and accurate prognostic information on smaller groups conditioned on follow-up time, stratified on age, stage and possibly other prognostic variables, will be challenging and expand the traditional cancer registry reports, probably beyond the comfort zone [99].

There are several large ongoing projects to make cancer patient survival estimates accessible to health care professionals and patients via online prognostic resources. These prediction tools aim to provide individual-level predictions of, e.g., metastatic progression, cause-specific survival and/or life expectancy, based on a profile of values of prognostic factors that are entered into the web-based application by the user. Predictions from elaborate statistical models are stored in a data base and

shown only for the patient profile that is asked for in a given search query. Examples of such work include the Cancer Survival Query System (CSQS) [100], developed by the Division of Cancer Control and Population Sciences at the National Cancer Institute, the prediction resource developed and published by the Memorial-Sloan Kettering Cancer Center [101], as well as PREDICT, an on-line prediction resource for breast cancer survival, developed by the National Institute for Health Research and the University of Cambridge [102]. Similar large scale programs do not yet exist in Sweden, but seem like the logical next step to assist physicians in risk counselling and to help their patients to better understand the prognosis and the risks and benefits of various treatment strategies.

10.3 Statistics targeted towards lay audiences

A reasonable question to ask is if we should be producing prognostic statistics aimed towards patients at all, given the impossible task to foresee the future of an individual patient. It must also be recognised that not all patients wish to speculate about their prognosis, and that physicians might (quite rightly so) be reluctant to quote a single number that serves as a best guess of the predicted prognosis. At the same time, information of this type is already communicated in the clinics, although in more subtle terms than actual percentages, and patients and their families are already known to actively search for information about their disease from web-based resources [19, 103, 104]. To meet their information needs, whether it is to assemble an information base for risk counselling, or to increase the general understanding of the potential implications of the disease and treatment for the affected patient/family member, we believe they should have access to the most recent and relevant data. We are, however, not suggesting that observational data and prognostic information from prediction tools of the type discussed above should be used in isolation, but rather as a complement to other scientific reports, and the clinical judgement and experience of the treating physician.

There is a vast amount of research and literature available on the topic of best practices for reporting medical information to lay people, see for example [16, 14, 105]. When it comes to presenting data, it is crucial to understand that patients are heterogeneous with respect to their information literacy, numeracy, and susceptibility to framing effects. In order to convey a clear message, all efforts to make data publicly available via e.g., web-based prediction tools must therefore be appropriately targeted and adapted for the intended end-user. For example, visual tools that use natural frequencies to display information have been found preferable to probability/percentage statements, or relative risk measures that are quoted without reference to the absolute risks [106]. Another issue that needs to be carefully explained when patients are seeking information outside the health care system, is the distinction between single event probabilities and probabilities applicable on a group level [107]. For this, we strongly advocate the use of disclaimers that explain the conditions for interpretation, and that a complete picture for the individual patient is best given by the patients own care provider.

11 Future perspectives

The opportunities to make more efficient use of the internationally acclaimed health registers that we have in Sweden are great. The Swedish cancer quality registers are starting to become mature, and to reach a high level of completeness. It seems timely to review what information physicians, who initiated and continuously report to these registers, wish to be able to access and bring into their daily clinical work. With this thesis, we hope to have given some new ideas of how population-based cancer register data could be used to investigate clinically relevant research questions, and to summarise cancer patient survival statistics. My hope is that the work described can inspire new work in the direction of making cancer patient survival statistics accessible to a broader range of users. In addition to further tailoring cancer patient survival statistics to meet the needs of the end users, and initiating the work towards developing a query system for presenting information on cancer patient survival, a couple of methodological issues that require additional research are described in this chapter.

11.1 Adjusting for comorbidities in studies of cancer patient survival

Relative survival is calculated as the ratio of the all-cause survival of the patients to the expected survival of a comparable cohort from the general population, assumed free from the disease under investigation. Apart from sex and calendar year, the expected survival is typically matched on each patient's chronological age. However, adjusting for these factors alone might not be sufficient to make the patients exchangeable to the general population. For example, among men with localised prostate cancer we found that those who were diagnosed with low risk disease had, on average, better all-cause survival than expected (i.e. negative excess mortality). One explanation for this phenomenon is that the "biological age" of men with localised prostate cancer might, in fact, be lower than their chronological age. The presence of comorbidities affects the biological age of individuals, yet comorbidities are very rarely considered in population-based studies of cancer patient survival. When comorbidities are considered, the Charlson comorbidity index [108] is typically used. The Charlson comorbidity index was developed to predict one-year all-cause mortality among patients enrolled in clinical trials and may therefore not be optimal for studying long-term cancer patient survival. An alternative approach to adjusting for comorbidities in cancer survival studies has been proposed by researchers at the National Cancer Institute, and involves matching patients to a reference population based also on the "biological"/health status adjusted age [100]. Determining an individual patient's health status adjusted age was done by using the number and type of comorbid conditions recorded in a national social insurance programme (Medicare). In this example the investigators chose to include 15 conditions that are also included in the Charlson comorbidity index, but in principle any combination of conditions can be included in the modelling of the health status adjusted age. In Sweden, it would be possible to use data recorded in the in- and out-patient hospital registers, but possibly also the Prescribed Drug Register, to identify the comorbid conditions that have the strongest influence on the patients' "biological" age.

A second approach to making cancer patients exchangeable to general population is to stratify the conventionally used population mortality files on additional factors. US life tables are, for example, often stratified on race, but suggestions and examples of population mortality files that include social class, geographic region and smoking status also exist [109, 110, 95, 111]. Irrespective of the preferred approach, to develop and apply methods that allow adjusting for additional factors which make patients heterogeneous from the general population is important to further improve predictions of survival in population-based cancer studies.

11.2 Assumptions for the expected survival in the cancer-free population

Standard errors of relative survival ratios are calculated without taking random variation in the expected survival into consideration. Variation in the expected survival has generally been viewed as negligible since the population life tables, from which the expected survival is typically retrieved, are derived from a large sample (the entire population). The expected survival in a given cohort of patients is, however, still subject to random variation resulting from random variation in the age distribution of the cancer patients. Moreover, the random variation in the observed and expected survival of cancer patients may be positively correlated, as both tend to decrease with increasing age. Brenner and Hakulinen demonstrated empirically that the latter correlation was of such magnitude that standard errors that accounted for the covariance between observed and expected survival were generally lower than those standard errors that are conventionally used (and which entirely ignore expected variation in relative survival calculations) [112]. In practice, this finding has not had an influence on how standard errors for relative survival survival are calculated. Reasons for this might be that the well-known Greenwood method at least gives conservative estimates of the standard error, and that bootstrap confidence intervals are found too computationally intensive in practice. In this thesis, we have also relied on the assumption that any random variation in the expected survival/mortality is negligible, and have treated them as fixed in the estimation and calculation of crude survival probabilities. A theoretical extension that is warranted involves assessing the validity of this assumption. In particular, such evaluation is important if there is a general movement in the field towards stratifying population life tables on additional factors, such as socio-economic status, race and/or comorbidity indices to satisfy the exchangeability assumption. Increasing the dimension of population life tables inevitably also makes the argument that the variability in the expected survival is negligible less convincing (since the data that gave rise to the life tables become more sparse). Seppä *et al.* have recently studied the impact of treating region-specific expected mortality rates as random quantities, as opposed to fixed, in the context of quantifying region-specific cure proportions among patients with colon cancer in Finland [113]. Only minor differences in the estimated cure proportions and median survival times of the uncured were seen when the expected mortality rates were treated as random. Studies that investigate the consequences of treating expected mortality rates as fixed in the settings of other cancer types and study designs are, however, still needed.

12 Acknowledgements

The research described in this thesis was carried out at the Department of Medical Epidemiology and Biostatistics between the years of 2010 and 2013. My time as a PhD candidate at Karolinska Institutet has been truly rewarding, academically as well as personally; for that I owe a debt of gratitude to many people who have inspired me along this journey.

- **Paul Dickman** - I feel very fortunate for having had you as my main supervisor. Thank you for so generously sharing your knowledge and research experience, wide network of collaborators, and teaching opportunities all around the world. I have appreciated your philosophy of what makes good doctoral education, and the fact that you have let me plan and take responsibility for my own education whilst, at the same time, always keeping your door open for questions, interesting discussions and support. In my world, this has been doctoral education at its best!
- **Paul Lambert** - Thank you for an excellent collaboration! It has been a joy having you as my co-supervisor, and I really appreciate all the time and support you have given me over these years. At times when I have entered unknown statistical or programming territory, you have always been one step ahead and made everything seem so doable and easy. Thank you also for inviting me to come and work in Leicester. I have had some really enjoyable and productive times over there.
- **Magnus Björkholm** - I am very glad that you agreed to be my co-supervisor. I still laugh when I think about that time when I had some questions about Swedish Hodgkin lymphoma treatment over the years and you sent me a package that included more or less every PhD-thesis of relevance on the subject (including your own). That was brilliant! I have really enjoyed our collaboration, and I will always be grateful that I was given the chance to work with you during these years. Thanks!
- **Therese Andersson** - I could not have asked for a better PhD companion. Thanks for all the discussions, questions, answers, laughs, coffees, lunches, dinners, trips all over the world and friendship over the years, in particular during the many weeks we spent at Opal Court in Leicester. It has been such a fun journey, the only sad thing about it is that it has now come to an end. Clearly, it's time to set new goals. How about *Docent* in 2017?
- My professional journey at MEB, from starting out as a newly graduated statistician in the Applied biostatistics group to writing this thesis has been a truly enjoyable one. I owe many thanks to **Henrik Grönberg**, **Nancy Pedersen**, **Hans-Olov Adami** and to the various constellations of the **Executive group** over the years for that. Thank you for your continuous efforts to make MEB an inspiring, creative and fun place to do research.
- Thanks to **Jan Adolfsson**, **Pär Stattin** and **Olof Akre** for co-authoring the first paper in this thesis with me and for welcoming me into the work group of the National Prostate Cancer Registry. Thanks also to **Per Hall** and **Kamila Czene** and **Janne Sjöberg** for all your help with my second and third studies, respectively.
- **Anna Johansson** and **Caroline Weibull** - I feel so lucky to have you as my close friends. Thank you for all the talk, laughs, support, opera, dinners, travel, and a gazillion other things we have done over the years. Thank you both also for letting me tag along to ParCa2-group meetings at times when I needed to put my thesis work to the side for a while.

- My Leicester colleagues **Mark Rutherford**, **Sally Hinchliffe** and **Michael Crowther** - Thanks for keeping an eye out for me and Therese in Leicester and for making sure that we had, not only a very productive work experience over in the curry capital of England, but also a busy social schedule.
- **Timo Hakulinen**, **Mark Clements**, **Mats Talbäck** and **Ronald Geskus**- I´ve really enjoyed our irregular discussions about research over the years even though my brain sometimes seemed incapable of processing information at the same speed as yours. I have learnt a lot from you and for that I am really grateful.
- A warm thanks to all the brilliant minds of **HEMEP** for great discussions, ideas, inspiration and support. I have genuinely enjoyed our collaboration and I hope it will continue in some form for many years to come.
- **Marie Jansson** - Thanks for looking after me and for reminding me to take lunch, to go home, and to hand in and sign Primula errands on time. What would we do without you in the corridor? Thanks also to **Camilla Ahlqvist** for helping me to keep track of my administrative journey towards becoming a PhD.
- Thanks to **Yudi Pawitan** and all **participants of the Likelihood course** for making it the best statistics course at KI. It was probably the most challenging course I ever took, but also one of the most fun learning experiences.
- **Silvelyn Zwanzig** - My former teacher and mentor. Thank you for raising my interest in statistical theory and for making challenging topics seem less frightening by mixing up all the mathematic formulae with a history lesson about famous statisticians.
- A big thanks to **Cecilia Lundholm** for making it possible for me to combine the first 2 years of my PhD-studies with suitable projects in the **Applied biostatistics group**. Thanks also to the rest of the **Biostatistics group** at MEB, in particular **Yudi Pawitan**, **Anna Johansson** and **Mark Clements** for your continuous efforts to make sure that we all stick together in a department that never seems to stop growing.
- To the past and present members of the **PUBMEB group**, especially **Caroline Weibull**, **Thomas Frisell** and **Robert Karlsson** who came along right from the very beginning. I never expected that a job-description at KI would ever include organising pubs, Christmas-parties and BBQs. Thanks for all the enthusiasm, optimism, hard work and fun times in Ljugården.
- **Susanna Toivanen** - Thanks for all your support (and for giving me a lovely place to live over the past three years). Our dinner pep-talks about life and work have helped me to figure out so much about what I want to do.
- Finally, thank you my dearest family and friends for your never-ending love and support and for helping me to keep all things in perspective.

References

- [1] Adami HO. Sweden: A paradise for epidemiologists? *Lancet* 1996;**347**:588–589.
- [2] Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000;**100**:57–70.
- [3] Cancer incidence in Sweden 2011 - Cancerförekomst i Sverige 2011 - [artikelnr: 2012-12-19]. Tech. rep., The National Board of Health and Welfare (Socialstyrelsen). Stockholm, Sweden, 2012.
- [4] The future population of Sweden 2012-2060. Tech. rep., Demographic reports 2012:2. Statistics Sweden (Statistiska Centralbyrån), 2012.
- [5] The Swedish Cancer Society (Cancerfonden). Vad orsakar cancer?, 2013. <http://www.cancerfonden.se> (Accessed 10 Oct 2013).
- [6] Ett halvt sekel med det svenska cancerregistret [artikelnr: 2008-126-48]. The National Board of Health and Welfare (Socialstyrelsen), Stockholm, Sweden, 2008.
- [7] The World Health Organisation (WHO). National cancer control programmes. Policies and Managerial Guidelines, 2nd edition. Tech. rep., Geneva, 2002.
- [8] Armstrong BK. The role of the cancer registry in cancer control. *Cancer Cause Control* 1992;**3**:569–579.
- [9] 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.
- [10] Parkin DM. The role of cancer registries in cancer control. *Int J Clin Oncol* 2008;**13**:102–111.
- [11] Dickman PW, Adami HO. Interpreting trends in cancer patient survival. *J Intern Med* 2006;**260**:103–117.
- [12] Johansson LA, Westerling R. Comparing Swedish hospital discharge records with death certificates: implications for mortality statistics. *Int J Epidemiol* 2000;**29**:495–502.
- [13] Bolund C, Brandberg Y. Psykosocial onkologi. In: Ringborg U, Dalianis T, Henriksson R, eds., *Onkologi*. Liber, 2008; 586–597.
- [14] Fagerlin A, Zikmund-Fisher BJ, Ubel PA. Helping patients decide: ten steps to better risk communication. *J Natl Cancer Inst* 2011;**103**:1436–1443.
- [15] Bolund C. Etik i sjukvården. In: Ringborg U, Dalianis T, Henriksson R, eds., *Onkologi*. Liber, 2008; 608–620.
- [16] Lipkus IM. Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. *Med Decis Making* 2007;**27**:696–713.

- [17] The World Internet Project. The fourth report of the World Internet Project, 2012. <http://www.worldinternetproject.net> (Accessed 15 July 2013).
- [18] Carlsson M. Cancer patients seeking information from sources outside the health care system. *Support Care Cancer* 2000;**8**:453–457.
- [19] Carlsson ME. Cancer patients seeking information from sources outside the health care system: change over a decade. *Eur J Oncol Nurs* 2009;**13**:304–305.
- [20] Lag (1998:543) om hälsodataregister (the Health Data Registers Act). Government Offices of Sweden, Ministry of Health and Social Affairs, Stockholm, Sweden 1998.
- [21] 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.
- [22] Van Hemelrijck M, Wigertz A, Sandin F, Garmo H, Hellström K, Fransson P, Widmark A, Lambe M, Adolfsson J, Varenhorst E, Johansson JE, Stattin P, *et al.*. Cohort profile: The National Prostate Cancer Register of Sweden and Prostate Cancer data Base Sweden 2.0. *Int J Epidemiol* 2012;.
- [23] Jorner U. *Summa Summarum SCBs första 150 år*. Statistics Sweden (Statistiska Centralbyrån), 2008.
- [24] Wilmoth JR, Andreev K, Jdanov D, Glej DA. Methods protocol for the Human Mortality Database. Technical report Version 5, 2007. Accessible from: <http://www.mortality.org>.
- [25] Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekblom A. The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *Eur J Epidemiol* 2009;**24**:659–667.
- [26] Berkson J. The calculation of survival rates. In: Walters W, Gray H, Priestly J, eds., *Carcinoma and other malignant lesions of the stomach*. Philadelphia: Sanders, 1942; 467–484.
- [27] Pohar Perme M, Stare J, Estève J. On estimation in relative survival. *Biometrics* 2012; **68**:113–120.
- [28] Rutherford MJ, Dickman PW, Lambert PC. Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiology* 2012;**36**:16–21.
- [29] 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.

- [30] Dickman PW, Lambert PC, Coviello E, Rutherford MJ. Estimating net survival in population-based cancer studies. *Int J Cancer* 2013;**133**:519–21.
- [31] Roche L, Danieli C, Belot A, Iwaz J, Remontet L, Bossard N. Author’s reply to: Estimating net survival in population-based cancer studies. *Int J Cancer* 2013;**133**:522–523.
- [32] Hakulinen T, Tenkanen L. Regression analyses of relative survival rates. *Applied Statistics* 1987;**36**:309–317.
- [33] 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.
- [34] 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.
- [35] 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. *Stat Med* 2003;**22**:2767–2784.
- [36] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Stat Med* 2004;**23**:51–64.
- [37] Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med* 2007;**26**:5486–5498.
- [38] Sasieni PD. Proportional excess hazards. *Biometrika* 1996;**83**:127–141.
- [39] Pohar Perme M, Henderson R, Stare J. An approach to estimation in relative survival regression. *Biostatistics* 2009;**10**:136–146.
- [40] Aalen O, Borgan O, Gjessing H. *Survival and Event History Analysis*. New York: Springer, 2008.
- [41] Cortese G, Scheike TH. Dynamic regression hazards models for relative survival. *Stat Med* 2008;**27**:3563–3584.
- [42] Zahl PH, Aalen OO. Adjusting and comparing survival curves by means of an additive risk model. *Lifetime Data Anal* 1998;**4**:149–168.
- [43] Stata. *Stata base reference manual Release 13*. Stata Press: College Station, 2013.
- [44] Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer* 1996;**78**:2004–2010.

- [45] Brenner H, Gefeller O, Hakulinen T. Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer* 2004;**40**:326–335.
- [46] Gondos A, Holleczeck B, Janssen-Heijnen M, Brewster DH, Bray F, Rosso S, Hakulinen T, Brenner H, E. Model-based projections for deriving up-to-date cancer survival estimates: an international evaluation. *Int J Cancer* 2009;**125**:2666–2672.
- [47] 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.
- [48] 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.
- [49] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007;**26**:2389–2430.
- [50] Klein J, Moeschberger ML. *Survival analysis: techniques for censored and truncated data*. Springer (Second edition), 2003.
- [51] Gail MH. Estimation and interpretation of models of absolute risk from epidemiologic data, including family-based studies. *Lifetime Data Anal* 2008;**14**:18–36.
- [52] Lambert PC, Dickman PW, Nelson CP, Royston P. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Stat Med* 2010;**29**:885 – 895.
- [53] Carstensen B. Demography and epidemiology: Practical use of the Lexis diagram in the computer age, or: Who needs the Cox-model anyway? Tech. rep., Steno Diabetes Center, Denmark, 2005.
- [54] Andersson TML, 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;**11**:96.
- [55] Friberg S. 5-year cure rate: yet another myth. *J Surg Oncol* 1997;**65**:73–75.
- [56] What is cure? Paths and Progress Fall/Winter 2008. Dana-Farber Cancer Institute. Vol 17, 2:8-10, 2008.
- [57] Boag J. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J R Stat Soc Ser B Stat Methodol* 1949;**11**:15–44.
- [58] De Angelis R, Capocaccia R, Hakulinen T, Söderman B, Verdecchia A. Mixture models for cancer survival analysis: Application to population-based data with covariates. *Stat Med* 1999;**18**:441–454.

- [59] Verdecchia A, Angelis RD, Capocaccia R, Sant M, Micheli A, Gatta G, Berrino F. The cure for colon cancer: results from the EURO CARE study. *Int J Cancer* 1998;**77**:322–329.
- [60] Yu B, Tiwari RC, Cronin KA, Feuer EJ. Cure fraction estimation from the mixture cure models for grouped survival data. *Stat Med* 2004;**23**:1733–1747.
- [61] 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.
- [62] 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.
- [63] 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; (In press).
- [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] Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. *Br J Cancer* 2004;**91**:1229–1235.
- [66] Berry SD, Ngo L, Samelson EJ, Kiel DP. Competing risk of death: an important consideration in studies of older adults. *J Am Geriatr Soc* 2010;**58**:783–787.
- [67] Bakoyannis G, Touloumi G. Practical methods for competing risks data: a review. *Stat Methods Med Res* 2012;**21**:257–272.
- [68] Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? *Stat Med* 2012;**31**:1089–1097.
- [69] Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol* 2012;.
- [70] Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. *Clin Cancer Res* 2012;**18**:2301–2308.
- [71] Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? *Nephrol Dial Transplant* 2013;.
- [72] Stattin P, Holmberg E, Johansson JE, Holmberg L, Adolfsson J, Hugosson J, on behalf of the National Prostate Cancer Register (NPCR) of Sweden. Outcomes in localized prostate cancer: National Prostate Cancer Register of Sweden follow-up study. *J Natl Cancer Inst* 2010;**102**:950–958.

- [73] Berglund A, Garmo H, Tishelman C, Holmberg L, Stattin P, Lambe M. Comorbidity, treatment and mortality: a population based cohort study of prostate cancer in PCBaSe Sweden. *J Urol* 2011;**185**:833–839.
- [74] Kim MM, Hoffman KE, Levy LB, Frank SJ, Pugh TJ, Choi S, Nguyen QN, McGuire SE, Lee AK, Kuban DA. Prostate cancer-specific mortality after definitive radiation therapy: who dies of disease? *Eur J Cancer* 2012;**48**:1664–1671.
- [75] Epstein MM, Edgren G, Rider JR, Mucci LA, Adami HO. Temporal trends in cause of death among Swedish and US men with prostate cancer. *J Natl Cancer Inst* 2012;**104**:1335–1342.
- [76] Daskivich TJ, Fan KH, Koyama T, Albertsen PC, Goodman M, Hamilton AS, Hoffman RM, Stanford JL, Stroup AM, Litwin MS, Penson DF. Effect of age, tumor risk, and comorbidity on competing risks for survival in a U.S. population-based cohort of men with prostate cancer. *Ann Intern Med* 2013;**158**:709–717.
- [77] Shao YH, Demissie K, Shih W, Mehta AR, Stein MN, Roberts CB, Dipaola RS, Lu-Yao GL. Contemporary risk profile of prostate cancer in the United States. *J Natl Cancer Inst* 2009; **101**:1280–1283.
- [78] Bratt O, Berglund A, Adolfsson J, Johansson JE, Törnblom M, Stattin P, for the Steering Committee of the Swedish National Prostate Cancer Register. Prostate cancer diagnosed after prostate-specific antigen testing of men without clinical signs of the disease: a population-based study from the National Prostate Cancer Register of Sweden. *Scand J Urol Nephrol* 2010;**44**:384–390.
- [79] Nordström T, Aly M, Clements MS, Weibull CE, Adolfsson J, Grönberg H. Prostate-specific antigen (PSA) testing is prevalent and increasing in Stockholm County, Sweden, despite no recommendations for PSA screening: results from a population-based study, 2003-2011. *Eur Urol* 2013;**63**:419–425.
- [80] Healey Bird BRJ, Swain SM. Cardiac toxicity in breast cancer survivors: review of potential cardiac problems. *Clin Cancer Res* 2008;**14**:14–24.
- [81] Darby SC, Ewertz M, McGale P, Bennet AM, Blom-Goldman U, Brønnum D, Correa C, Cutter D, Gagliardi G, Gigante B, Jensen MB, Nisbet A, *et al.* Risk of ischemic heart disease in women after radiotherapy for breast cancer. *N Engl J Med* 2013;**368**:987–998.
- [82] Andersen PK, Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res* 2002;**11**:91–115.
- [83] Rutqvist LE, Johansson H. Mortality by laterality of the primary tumour among 55,000 breast cancer patients from the Swedish Cancer Registry. *Br J Cancer* 1990;**61**:866–868.

- [84] Darby S, McGale P, Peto R, Granath F, Hall P, Ekbom A. Mortality from cardiovascular disease more than 10 years after radiotherapy for breast cancer: nationwide cohort study of 90 000 Swedish women. *BMJ* 2003;**326**:256–257.
- [85] Darby SC, McGale P, Taylor CW, Peto R. Long-term mortality from heart disease and lung cancer after radiotherapy for early breast cancer: prospective cohort study of about 300,000 women in US SEER cancer registries. *Lancet Oncol* 2005;**6**:557–565.
- [86] Fang F, Fall K, Mittleman MA, Sparén P, Ye W, Adami HO, Valdimarsdóttir U. Suicide and cardiovascular death after a cancer diagnosis. *N Engl J Med* 2012;**366**:1310–1318.
- [87] Sjöberg J, Halthur C, Kristinsson SY, Landgren O, Nygell UA, Björkholm M, Dickman PW. Progress in Hodgkin lymphoma: a population-based study on patients diagnosed in Sweden from 1973-2009. *Blood* 2012;**119**:990–996.
- [88] Glimelius B, Enblad G, Kalkner M, Gustavsson A, Jakobsson M, Branehog I, Lenner P, Bjorkholm M. Treatment of Hodgkin’s disease: the Swedish National Care Programme experience. *Leuk Lymphoma* 1996;**21**:71–78.
- [89] Andersson A, Näslund U, Tavelin B, Enblad G, Gustavsson A, Malmer B. Long-term risk of cardiovascular disease in Hodgkin lymphoma survivors—retrospective cohort analyses and a concept for prospective intervention. *Int J Cancer* 2009;**124**:1914–1917.
- [90] Ng AK, Bernardo MP, Weller E, Backstrand KH, Silver B, Marcus KC, Tarbell NJ, Friedberg J, Canellos GP, Mauch PM. Long-term survival and competing causes of death in patients with early-stage Hodgkin’s disease treated at age 50 or younger. *J Clin Oncol* 2002;**20**:2101–2108.
- [91] Aleman BMP, van Leeuwen FE. Are we improving the long-term burden of Hodgkin’s lymphoma patients with modern treatment? *Hematol Oncol Clin North Am* 2007;**21**:961–975.
- [92] Baxi SS, Matasar MJ. State-of-the-art issues in Hodgkin’s lymphoma survivorship. *Curr Oncol Rep* 2010;**12**:366–373.
- [93] Lee R, Carter L. Modeling and forecasting U.S. mortality rates. *J Am Stat Assoc* 1992;**87**:659–671.
- [94] Yu B, Tiwari RC, Feuer EJ. Estimating the personal cure rate of cancer patients using population-based grouped cancer survival data. *Stat Methods Med Res* 2010;.
- [95] 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.
- [96] Andersson TML, 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;**148**:918–924.

- [97] Andersson TML, Lambert PC, Eriksson H, Månsson-Brahme E, Hansson J, Dickman PW, Eloranta S, Lambe M. Estimating the cure proportion of malignant melanoma, a population-based study in 2 regions of Sweden. Oslo, Norway, 2012. The Annual meeting of the Association of the Nordic Cancer Registries, 2012; Conference abstract.
- [98] Baade PD, Youlten DR, Chambers SK. When do I know I am cured? Using conditional estimates to provide better information about cancer survival prospects. *Med J Aust* 2011; **194**:73–77.
- [99] Møller B, Aagnes B. Conditional survival statistics in standard cancer registry reporting. Oslo, Norway, 2012. The Annual meeting of the Association of the Nordic Cancer Registries, 2012; Conference abstract.
- [100] Feuer EJ, Lee M, Mariotto AB, Cronin KA, Scoppa S, Penson DF, Hachey M, Cynkin L, Carter GA, Campbell D, Percy-Laurry A, Zou Z, *et al.*. The Cancer Survival Query System: making survival estimates from the Surveillance, Epidemiology, and End Results program more timely and relevant for recently diagnosed patients. *Cancer* 2012;**118**:5652–5662.
- [101] Ross PL, Scardino PT, Kattan MW. A catalog of prostate cancer nomograms. *J Urol* 2001; **165**:1562–1568.
- [102] Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, Caldas C, Pharoah PD. PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010;**12**:R1.
- [103] Nilsson-Ihrfelt E, Fjällskog ML, Blomqvist C, Ahlgren J, Edlund P, Hansen J, Malmberg L, Villman K, Andersson G. Breast cancer on the internet: the quality of Swedish breast cancer websites. *Breast* 2004;**13**:376–382.
- [104] Buntrock S, Hopfgarten T, Adolfsson J, Onelöv E, Steineck G. The internet and prostate cancer patients: searching for and finding information. *Scand J Urol Nephrol* 2007;**41**:367–374.
- [105] Neuner-Jehle S, Wegwarth O, Steurer J. [is one picture worth a thousand words? risk communication in primary care: advantages and shortcomings of different methods]. *Praxis Bern* 1994 2008;**97**:669–676.
- [106] Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. *BMJ* 2003;**327**:741–744.
- [107] Gigerenzer G, Galesic M. Why do single event probabilities confuse patients? *BMJ* 2012; **344**:e245.

- [108] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;**40**:373–383.
- [109] Dickman PW, Auvinen A, Voutilainen ET, Hakulinen T. Measuring social class differences in cancer patient survival: is it necessary to control for social class differences in general population mortality? A Finnish population-based study. *J Epidemiol Community Health* 1998;**52**:727–734.
- [110] Woods LM, Racht B, Riga M, Stone N, Shah A, Coleman MP. Geographical variation in life expectancy at birth in England and Wales is largely explained by deprivation. *J Epidemiol Community Health* 2005;**59**:115–120.
- [111] Blakely T, Soeberg M, Carter K, Costilla R, Atkinson J, Sarfati D. Bias in relative survival methods when using incorrect life-tables: lung and bladder cancer by smoking status and ethnicity in New Zealand. *Int J Cancer* 2012;**131**:E974–E982.
- [112] Brenner H, Hakulinen T. Substantial overestimation of standard errors of relative survival rates of cancer patients. *Am J Epidemiol* 2005;**161**:781–786.
- [113] Seppä K, Hakulinen T, Läärä E. Regional variation in relative survival—quantifying the effects of the competing risks of death by using a cure fraction model with random effects. *J R Stat Soc Ser C Appl Stat* 2013;(In press).