Long-term net survival among women diagnosed with breast cancer: accuracy of its estimation and evaluation of its determinants

ROBIN SCHAFFAR

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

March 2018

Department of Non-Communicable Disease Epidemiology

Faculty of Epidemiology and Population Health

London School of Hygiene and Tropical Medicine

Declaration of authorship

I, Robin Schaffar, confirm that the work presented in this thesis is my own. When information has been derived from other sources, I confirm that this has been indicated in the thesis.

Name: Robin Schaffar

Date: 25 Janvier 2018

Signature 

Use of published work

This is a research paper style thesis.

Three papers have been published during the period of PhD registration and are included within this thesis. A further paper is under review and is also included in its submitted form. Robin Schaffar was the lead and corresponding author on all four, planned and carried out the literature reviews and data analysis, and prepared all drafts of each paper. The co-authors worked with Robin to plan the content of each paper, provided input and feedback on the data analysis and comments on the drafts prepared by Robin Schaffar.

Supervisors

Dr Laura Woods, London School of Hygiene and Tropical Medicine.

Dr Bernard Rachet, London School of Hygiene and Tropical Medicine.

Advisory panel

Dr Aurélien Belot, London School of Hygiene and Tropical Medicine.

Dr Elisabetta Rapiti, Geneva Cancer Registry.

Examiners

Professor Sabine Siesling, Comprehensive Cancer Centre, Groningen, The Netherlands.

Professor Stephen Duffy, Centre for Cancer Prevention, London, United Kingdom.

Abstract

Breast cancer is a major public health challenge. It affects a very large numbers of women across the globe. Although improvements in its management have dramatically transformed its prognosis at diagnosis, breast cancer remains associated with an increased long-term risk of death, persisting even decades after diagnosis. A comprehensive understanding of this underlying pattern of death from breast cancer in the long-term is currently lacking but increasingly important as the number of long-term survivors rises. The reliability of the cause of death is of particular interest in this context.

In this thesis, I use data from the Geneva Cancer Registry to first, determine the best methodology for examining long-term net survival, and second, to evaluate its determinants.

Two data settings are available for the estimation of net survival: the cause-specific setting, where the cause of death is required, and the relative-survival setting, where it is not. I first evaluated the accuracy of routinely collected cause of death information and the impact of inaccuracies upon survival estimates. I observed small but non-negligible advantages in using a reviewed cause of death when estimating survival. I then compared the cause-specific to the relative survival setting for the estimation of long-term net survival and demonstrated that the relative-survival setting was less sensitive to violations of the assumptions both for breast cancer patients as well as for patients diagnosed with cancer at three other localisations.

I further investigated the long-term effects of key prognostic factors and treatment for women with breast cancer in the relative survival setting using an appropriate strategy for model selection. Although I demonstrated insightful non-linear and time-dependent effects for some prognostic variables, the analyses were limited by issues of convergence and misspecification of the model. High quality population-based data and additional statistical tools are required to understand with greater certainty the determinants of breast cancer long-term excess mortality.

Acknowledgments

Getting a PhD represents a milestone in a researcher life. I embarked on the adventure a few years ago and finally crossing the finish line represents to me the greatest achievement. It has been a long and winding road, sometimes very rocky. The people I wish to thank now will understand what I am talking about.

I have to start with Laura. I thank you very much for all your indefectible support along this long experience. I have been very fortunate to have you as my main supervisor. Words cannot express how grateful I am. You have been motivating, encouraging, and enlightening. You never judged nor pushed when I needed to juggle priorities. You have been a tremendous mentor for me. As was Bernard, my second supervisor. I would like to express my sincere gratitude for your patience, availability, and immense knowledge. I thank you both for encouraging my research and for allowing me to grow as a research scientist. All your advices, brilliant comments and suggestions have been priceless. I really hope that we will be able to stay in touch in the future.

Besides my supervisors, I would like to thank Aurélien, who jumped on the bandwagon and provided me with precious support. I thank you for your patience, flexibility, genuine caring and concern. I could not have imagined having a better advisor for all the statistical work and whatnot.

I must say a word about the fantastic team that is the Cancer Survival Group. I met amazing and incredibly smart people over there. It has always been, and will remain, a great pleasure to visit you in London. A special thanks to Yuki. She was one of the first friendly faces to greet me when I began this doctoral

program and has always been a tremendous help no matter the task or circumstance. My gratitude is also extended to Michel Coleman who warmly welcomed me into the team. I am grateful for his wisdom and infinite knowledge that he is always happy to share. He has, at a distance, always found the right words.

From London, I'm now flying to Geneva. All of this project would have been impossible without the Geneva Cancer Registry and its extraordinary people. I am indebted to my many of all of them to support me in a number of ways. I am now looking forward to be fully involved and stand with them again. Of course, I owe my deepest gratitude to Christine Bouchardy. None of this would have happened without you. You offered me the freedom of achieving this thesis and always believed in me. I just can't thank you enough.

I also have to thank my astoundingly supportive band members, Elisa, Massimo, Gerald who lived almost every single minute of it. I was spared from so many difficulties thanks to you. For all of this and your unflawed support, I am eternally grateful.

I have a special thought for Jean-Michel Lutz, who guided me towards the London School of Hygiene and Tropical Medicine in the first place.

These acknowledgments would not be complete without mentioning my close relatives. I would like to say a heartfelt thank you to my Mum and Dad, always believing in me and heartening me blindly to follow my dreams. They always encouraged me to strive towards my goal. I thank also, Martin and Margot, aka the M&M's, for their support and offering me the honour of being the godfather of Théo.

The best outcome from these past six years has been finding my best friend, soul-mate, and partner in crime. Clarisse, I thank you for all of the sacrifices that you've made on my behalf. These past several years have not been an easy ride, both academically and personally. I truly thank you for sticking by my side, even when I was irritable and depressed. The result of this was the greatest present on earth: Baby Seppi, to whom I am dedicating this thesis. He has been a twinkle in my eye since he was born. He is now the next big project of my life!

Funding

The PhD was partially funded by a bursary from the Swiss Cancer League [BIL KFS-3274-08-2013] and supported by the Geneva Cancer Registry.

Table of Contents

List of tables and figures

Oral presentations

- The first two papers entitled "Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva cancer registry" and "Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis" were presented as oral communication at the GRELL (Group of Registry and Epidemiology in Latin Language Countries) meeting in Geneva, May 2014.
- The second paper entitled "Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis" was presented as an oral communication at the Division of Clinical Epidemiology seminar, Geneva University Hospitals, May 2016.

Poster presentations

- The second paper entitled "Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis" was presented as a poster at the International Association of Cancer Registries Congress in Ottawa, June 2014.

Award

- The second paper entitled "Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis", published in Cancer Epidemiology, was awarded the 'Enrico Anglesio' Prize, offered by the 'Anglesio Moroni Foundation', Turin, Italy. The prize was received at the GRELL meeting in Geneva, May 2014.

[Research question, aims and objectives]

What is the most accurate way to estimate net survival when reviewed cause of death is available?

&

When using an accurate approach, what can be determined about the long-term effects of prognostic factors and treatment for women with breast cancer?

Aims of the thesis

The overall aims of this thesis are:

- to evaluate the accuracy of routine cause of death information collected within a cancer registry setting and the impact of inaccuracies on survival estimates,
- to evaluate the less biased data setting (cause-specific vs. relative-survival) for the estimation of long-term net survival when reviewed cause of death is available and
- to use the superior approach found in previous aim to investigate the long-term effects of key prognostic factors and treatment for women with breast cancer.

Objectives of the thesis

The specific objectives of the thesis follow the aims and are as follows

1. Evaluate data quality

    a. to investigate how accurate, the routinely recorded cause of death field is compared to the cause of death field derived from comprehensive clerical review and
    b. to compare cause-specific survival estimates of net survival using routinely recorded cause of death to those derived using cause of death clerically reviewed by trained registrars.

2. Evaluate the more accurate data setting (cause-specific vs. relative-survival) for the estimation of long-term net survival

    a. to derive up-to-date life tables for use in the estimation of net survival within the relative survival setting,
    b. to apply an inverse probability weighting (IPW) method to estimate net survival in the cause-specific setting
    c. to compare and contrast estimates of long-term net survival from breast cancer in each data setting using both routinely collected and validated data on cause of death and
    d. to assess whether these same findings apply to other anatomic localisations.

3. Apply the approach determined to be less biased in 2. to investigate the long-term effects of prognostic factors

    a. to model long-term excess mortality including clinical variables to evaluate the long-term effect of prognostic factors and
    b. to assess non-linear and time-varying effects of these factors using the most appropriate methods.

[BACKGROUND]

An important burden

Breast cancer is a major worldwide public health concern. It is the second most common cancer in the world after lung and, by far, the most frequent cancer among women. Globally, there are more than 6 million female breast cancer survivors [1]. Every 19 seconds, a women is diagnosed with breast cancer [2,3], and every 74 seconds, a woman dies from her breast cancer [3].

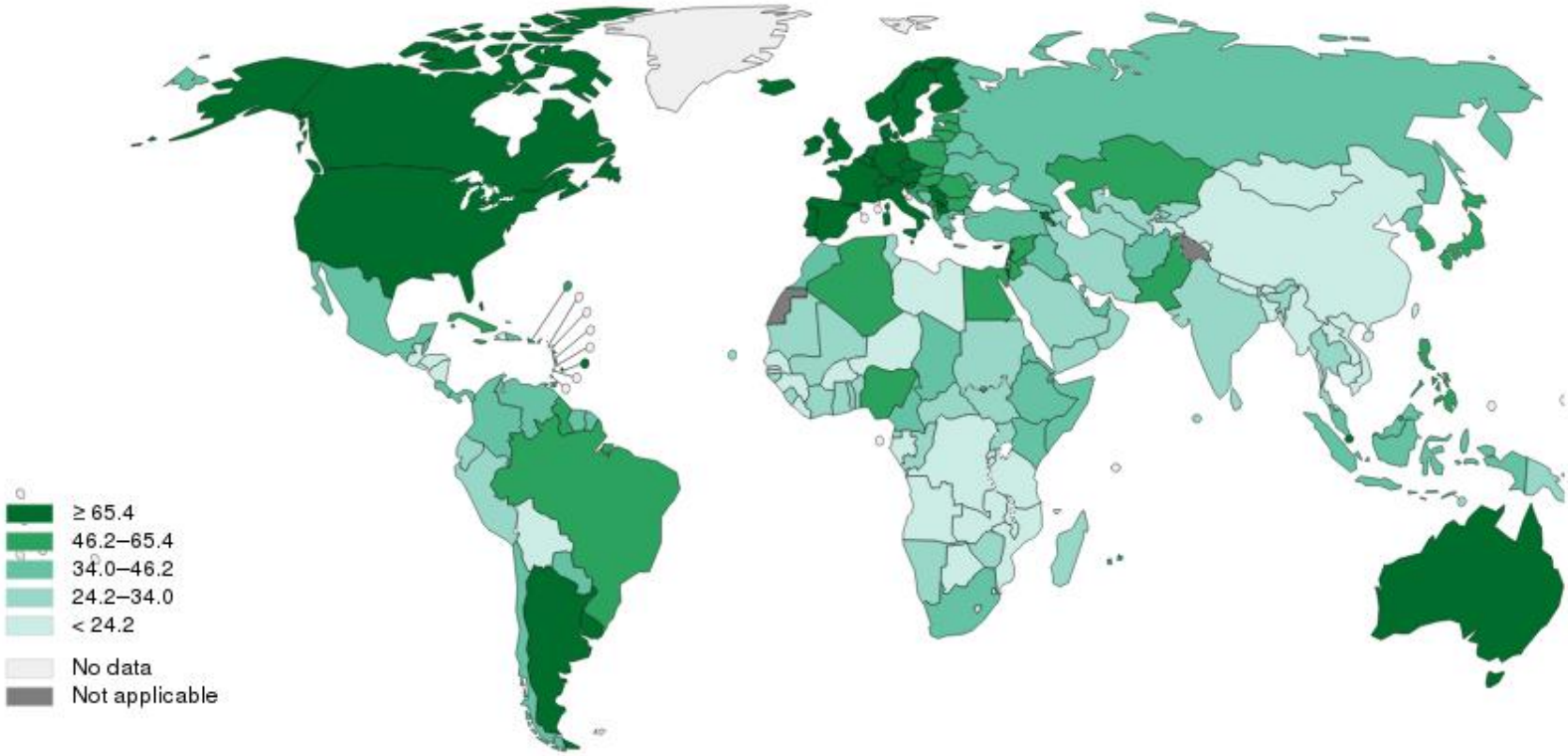
In 2012, globally there were:

- ✓ 1.7 million women newly diagnosed with breast cancer which represented 25% of all cancers and
- ✓ 522 000 deaths due to breast cancer.

Since 2008, breast cancer incidence around the world has increased by more than 20 percent, while mortality has increased by 14 percent [1]. At the current rate of increase, 41 million new cases of breast cancer are expected and 10.6 million women will die from breast cancer during the next 25 years worldwide [3].

The incidence rate of breast cancer varies considerably by global region (Figure 1). Since risk factors are mostly related to a more westernised lifestyle, incidence tends to be lower in developing countries but breast cancer remains the most common tumour among women. In 2008, the age-standardised incidence rate was 20 cases per 100,000 person-years in East Africa but reached 89.7 cases per 100,000 person-years in Western Europe.

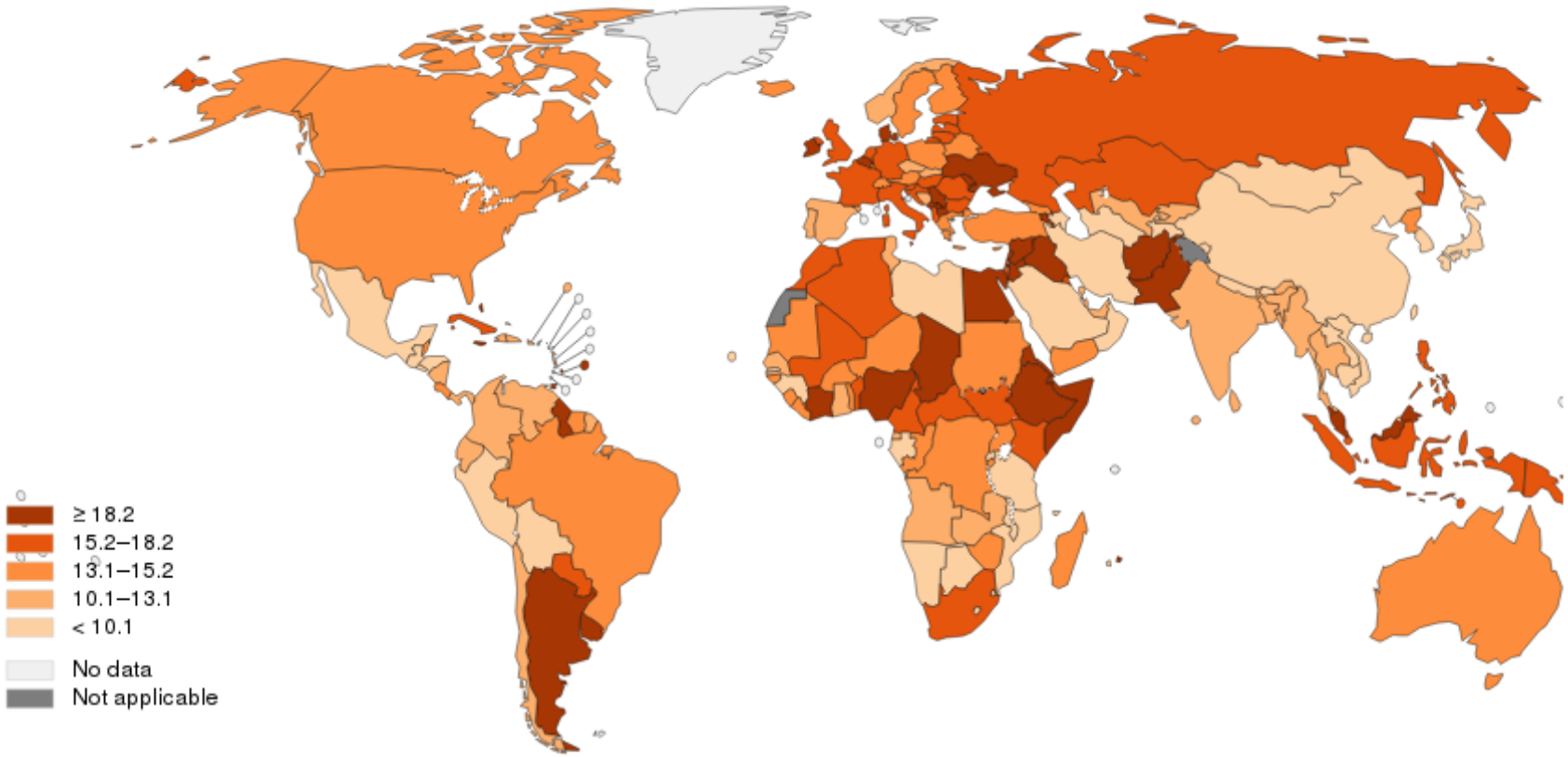
The range in mortality rates between world regions is less than that for incidence because survival tends to be higher in high-incidence, developed regions (Figure 2). For example, in 2012, the age-standardised mortality rate was 16.4 per 100,000 in France, against 23.0 per 100,000 in Ethiopia.



Data source: GLOBOCAN 2012
Map production: IARC
(http://gco.iarc.fr/today)
World Health Organization

World Health Organization

Figure 1: Estimated age-standardised incidence rates (per 100,000) female breast cancer, worldwide, 2012. Source: GLOBOCAN 2012.



Data source: GLOBOCAN 2012
Map production: IARC
(http://gco.iarc.fr/today)
World Health Organization

World Health Organization

Figure 2: Estimated age-standardised mortality rates (per 100,000), female breast cancer, worldwide, 2012. Source: GLOBOCAN 2012.

An increasing number of survivors

Despite the spread of breast cancer worldwide, the prognosis of the disease has dramatically improved over the last four decades. The scientific literature testifies to this by the number of studies demonstrating survival improvements [4–7]. The underlying explanation for this improvement in disease detection is both through screening programs and better treatment strategies. Surgical techniques have been improved, and new chemotherapies and hormonal agents have been developed to better fight the disease [8].

As a consequence, the number of women alive who have presented a diagnosis of breast cancer is increasing. This statement wouldn't be an issue if all these so-called survivors would be considered as definitely cured. This is, however, not so straightforward. Indeed, several studies have demonstrated that no 'cure' can be established at a population level. [9–18]. This does not mean to say that no individual patients can be cured but that there is a persistent excess hazard is associated with being diagnosed with breast cancer. This might be related to time-varying influencing factors and/or side effects following their disease [19]. We distinguish fatigue, difficulties with life insurance or return to work, lack of concentration but also more severe events like cardiac disease, bones issues, second primary malignancies and thromboembolic events, all being related to breast cancer and likely to cause death. Some of these sides effects are related to the breast cancer treatment and therefore mislead the allocation of the cause of death.

Patients who have survived a long time since their diagnosis thus represent a relatively new challenge in terms of public health, and there is a particular

interest in understanding their pattern of long-term mortality [20]. In this context, long-term survival is the key indicator.

Examining long-term survival represents a new challenge insofar as the implications of surviving after a breast cancer diagnosis beyond 5 or 10 years has not been widely considered. Indeed, one of the key assumptions of most survival models is the proportionality of the hazards along follow-up time, i.e. that the relative effects of the covariates included in the model remain the same throughout time after diagnosis [21]. However, what is true one year after diagnosis might be quite different later on. Several examples of time-varying effects are available in the literature [20–26]. For instance, in a cohort Norwegian cancer patients, Zahl [27] showed that the effect of stage sharply changed over time with a significant impact between 15 and 20 years. Jatoi *et al.* [24] shed the light on a time-dependent (non-proportional) hazard for breast cancer and showed that, for instance, hormone receptor status had a variable effects on survival over time after diagnosis.

Hence, being able to derive accurate long-term survival related to the disease and evaluate its long-term determinants by allowing them to vary through time since diagnosis represents an important first step towards the understanding of the pattern in long-term breast cancer mortality.

BOX 1

Summary

- Breast cancer has a significant and increasing burden worldwide
- The number of survivors is increasing, but 'cure' may not be reached
- Understanding the pattern of long-term mortality is important

What is next?

- What is breast cancer?

What is breast cancer?

Carcinogenesis

The human body is divided into several complex systems. A basic element of all these systems is the cell. Almost all cells are capable of division. Mutations can however occur because of an error in the DNA duplication during the cell division. These are caused by genetic instabilities, which are a result of exposure to carcinogens. This leads to an over-proliferation of cells and to uncontrolled growth. Such 'new growths' are called neoplasms or cancers. In contrast to benign cancers, malignant tumours spread and invade (metastasize) other organs in the body [28].

When the process of over-proliferation of the cells outside the normal control mechanisms of the body takes place within the breast, we have a breast cancer. Breast is defined as the primary site of the tumour.

The earliest evidence of cancer has been found in fossilized bone tumours and human mummies in ancient Egypt. The Edwin Smith Papyrus is the first written record of cancer (Figure 3) [29]. It was discovered about 3,000 BC in Egypt and described, amongst others, tumours of the breast and of their cauterisation [30].

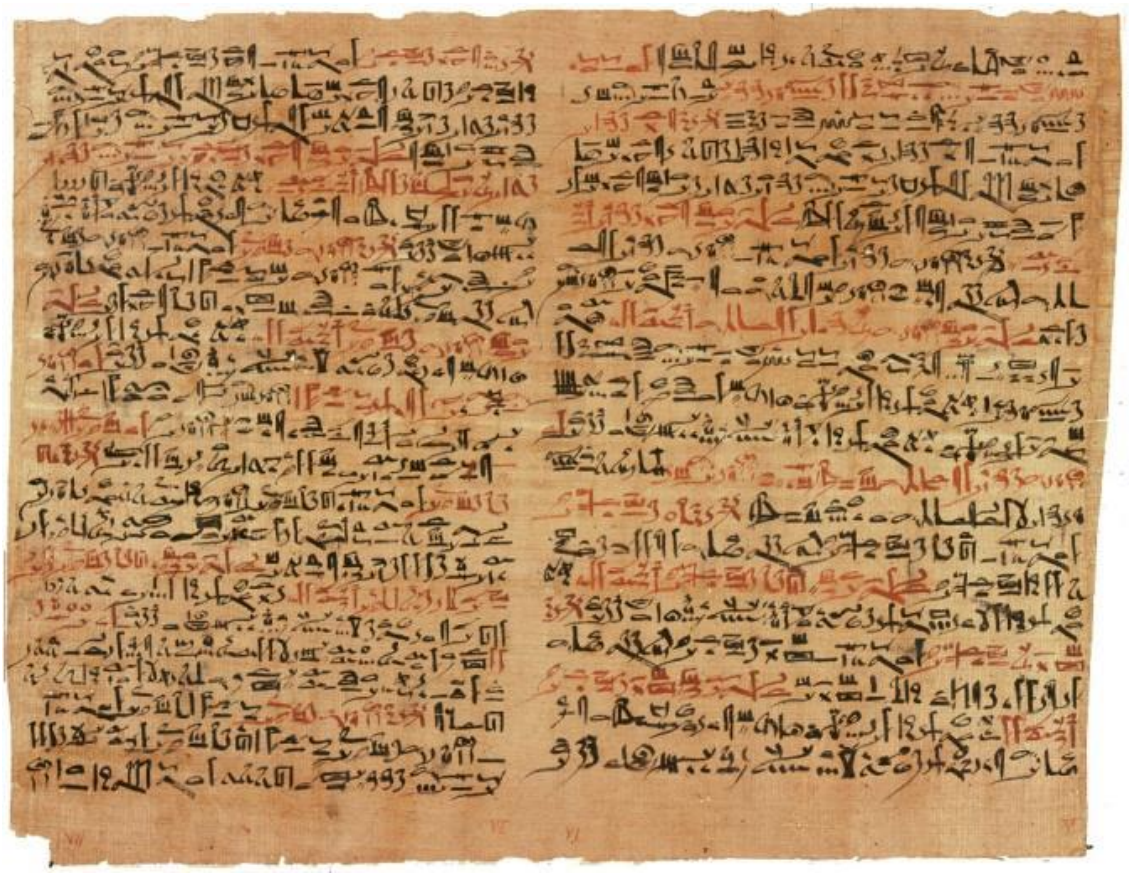


Figure 3: Edwin Smith Papyrus, found in Egypt about 3000 BC. Source: U.S. National Library of Medicine.

Breast anatomy

The breast is part of the reproductive system and is designed to produce milk. Milk is produced in the lobules (1) and goes to the nipples (2) through the lactiferous ducts (3). This milk production system is surrounded with adipose tissue (4) forming the breast (Figure 4) [31].

Breast tissue is epithelial, which is one of the four basic types of tissue found in the human body. Tumours growing within the epithelium are called carcinomas. These are the most common histological type among breast cancer. Invasive breast carcinomas consist of several histologic subtypes that differ with regard to their clinical presentation, radiographic characteristics, pathologic features, and biologic behaviour. The most common type of

invasive breast cancer are infiltrating ductal carcinoma accounting for 70 to 80 percent of invasive lesions and lobular carcinomas representing 10-15% of the breast tumours.[32]

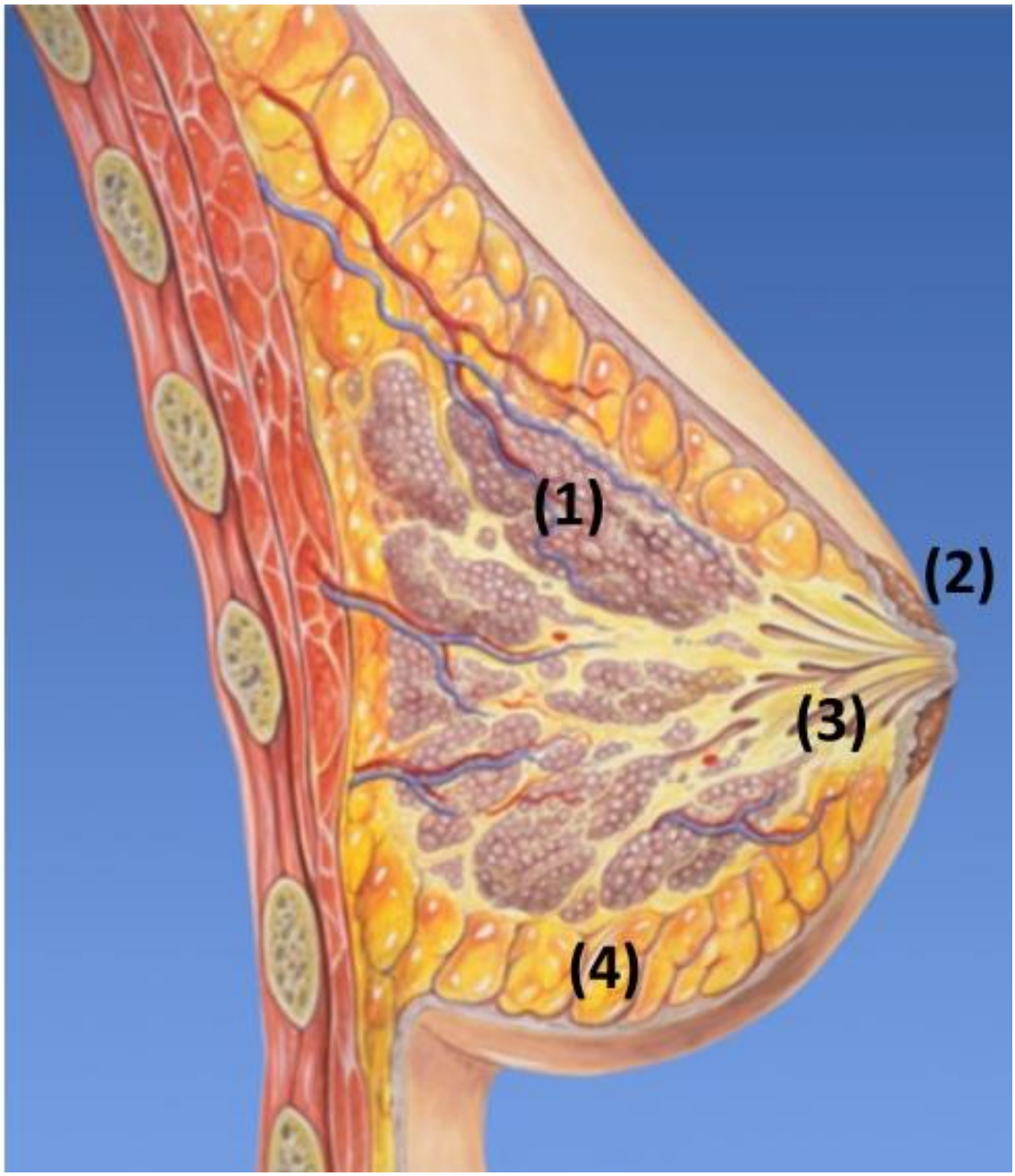


Figure 4: Breast anatomy. Source: Medical Illustrations by Patrick Lynch, generated for multimedia teaching projects by the Yale University School of Medicine, Centre for Advanced Instructional Media, 1987-2000.

The breast has very rich lymphatic drainage. The lymphatic system plays an important role in detoxification, immune response and hormone circulation. The lymphatic system is a circulatory system in the body and is divided into several parts. Lymph nodes are part of the immune system and are connected by lymphatic vessels. Most lymphatic vessels in the breast connect to lymph nodes under the arm (the axillary nodes) but they also connect to lymph nodes inside the chest (internal mammary nodes) and either above or below the collarbone (supra-clavicular or intra-clavicular nodes) (Figure 5).

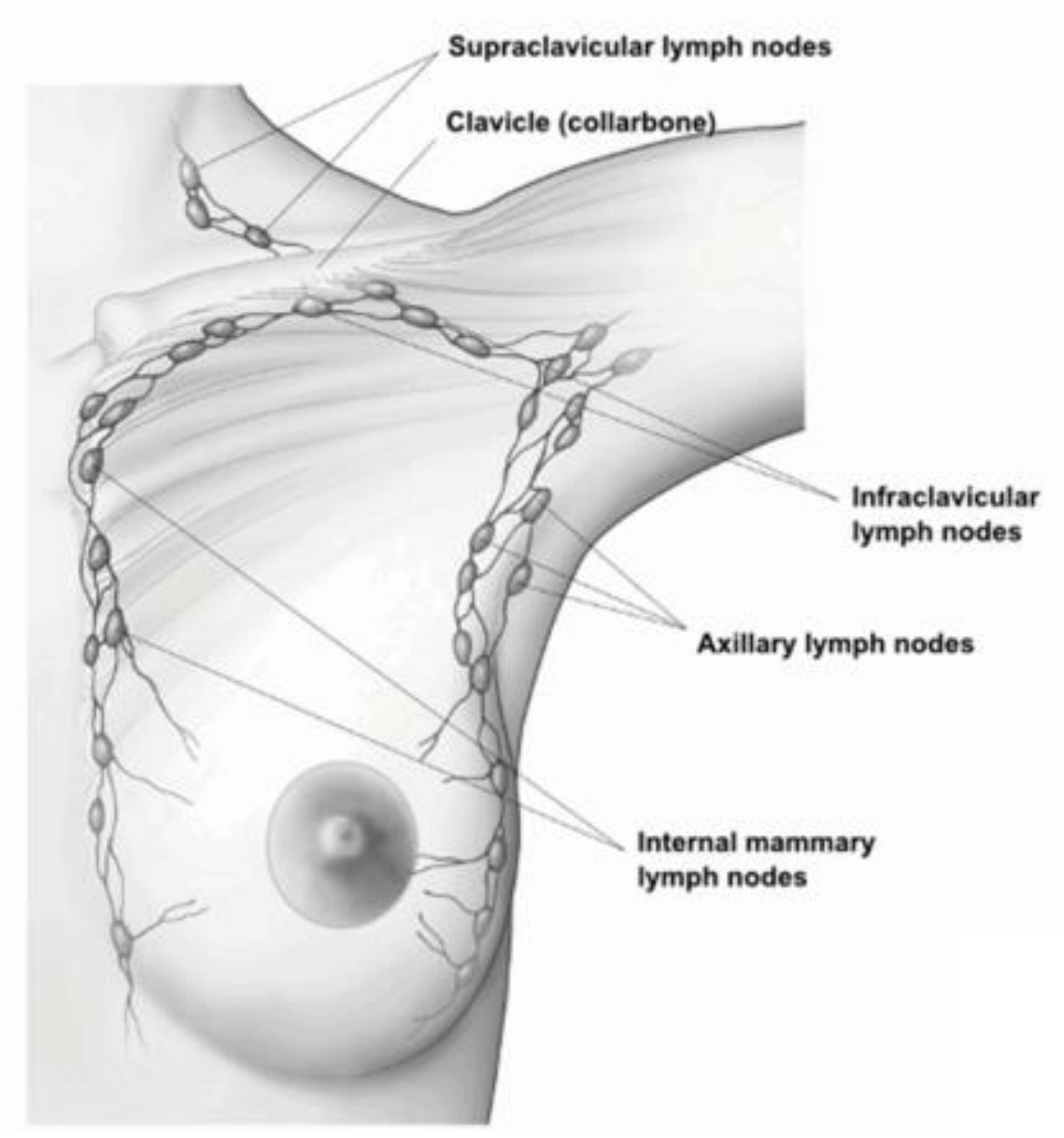


Figure 5: Lymph nodes in relation to the breast. Source: http://www.cancer.org/.

Cancer tumours have the ability to spread via the lymph nodes and the bloodstream to other sites in the body. Once the cancer enters the lymph nodes the more likely it is that the cancer may be found in other organs of the body. Because of this, identification of cancer in one or more lymph nodes often affects the treatment plan, since the spread of the disease is an indicator of its extension and of prognosis. We distinguish, in order of risk of death, localized, regional or distant disease. Treatment is planned according to this staging categorisation, from aggressive treatments with curative intent to palliative treatment. Although the degree of spread is closely linked to lymphatic involvement, not all women with cancer cells in their lymph nodes develop metastases, and some women with no cancer cells in their lymph nodes do develop later metastases. Bigger tumours are more susceptible to develop metastases. Breast cancers spread in particular to the bones, brain, liver and lungs [33].

Diagnosis

Several different symptoms are indicative of breast cancer. Principally these are lumps in the breast, an inverted nipple, distortion of skin or skin inflammation.

A number of investigations can be performed in order to diagnose breast cancer. Mammography, then biopsy followed by cytology and/or histology are all used to assess whether a lump is a malignant or benign breast cancer. The disease can also be asymptomatic for a period of time. Screening programs have been introduced to try to detect tumours as early as possible. Despite some remaining difficulties with dense breast tissue in young women, screening is generally accepted to be an effective method of reducing cancer mortality

[34,35]. The underlying assumption of screening is that detecting breast tumours early leads to better prognosis: more specifically, identifying tumours when they are smaller, and thus more likely to be at an early stage, reduces the risk that the disease has already spread. Further investigations upon diagnosis, such as bone scans or checking lymph nodes, allow the full spread of the disease to be established.

Causes and risks factors

There is not one cause of breast cancer but many predisposing factors [36]. Oestrogen and progesterone, related to the reproductive cycle, play an important role within the breast and the degree of exposure to these hormones is closely linked to the development of breast cancer tumours. All factors related to hormone regulation such as early menarche, late menopause, nulliparity, older age at first birth, absence of breast feeding, use of contraceptive pills or hormone replacement therapies increase the probability of breast cancer. Besides family history of breast cancer, genetic predisposition (BRCA 1 and BRCA 2), and personal history of benign breast disease are risk-factors for pre-menopausal disease. Obesity, leading to hormone and growth factor storage, as well as exposure to radiation, are also known risks factors.

Staging and grading

The prognosis for an individual woman is strongly determined by the tumour's characteristics. Stage is used to describe the degree of spread of the tumour at diagnosis. The most commonly used coding system for breast cancer is the Classification of Malignant Tumours also called TNM devised by the Union for International Cancer Control (UICC) and the American Joint Committee on

Cancer (AJCC) [36]. The smaller and more localised the tumour the better the prognosis of the disease.

Grade is about the evaluation of the level of differentiation of the tumour cells. It is a measure of the cell appearance and is based on the resemblance of the tumour to the tissue of origin [38]. A well-differentiated tumour still closely resembles the tissue of origin, whereas a poorly-differentiated tumour does not. From well differentiated to poorly differentiated, the grade of the tumour provides a measure of the capacity of the tumour to spread. Poorly-differentiated tumours have a greater capacity to spread compared with well-differentiated tumours.

Both staging and grading allow for decisions to be made about the management of patients and to identify appropriate treatment plans.

Prognostic factors

A prognostic factor is defined as a feature of the patient or cancer, which is measurable at the time of diagnosis that is associated with the outcome. The literature has described a large number of prognostic factors for breast cancer. We distinguish several types: those related to the patient herself, to the way the cancer is diagnosed, and to features of the tumour.

Patient characteristics

Age at diagnosis

Both younger and older ages at diagnosis are associated with a worse prognosis [38]. Tumours tend to be biologically more aggressive in younger patients [40], whilst older patients tend to be diagnosed later, have greater numbers of co-morbidities and be treated less aggressively [41].

Socio-economic deprivation

There is evidence that socio-economic deprivation is associated with poorer outcomes in many countries [42]. Women with higher socio-economic level are at higher risk of developing a breast cancer, but once the disease has occurred, their prognosis is better than for patients with a low socio-economic level. This may be explained by less aggressive tumours, more effective treatment strategies, comorbidities and lifestyle factors [42].

Co-morbidities

Patients with comorbidities have a higher risk of dying from their disease [43]. Comorbidities are usually categorised with the Charlson index, which consider 19 conditions, each weighted according to its potential to influence mortality [44].

Lifestyle status

The woman's general health condition plays a significant role in breast cancer survival [45]. As an illustration, low levels of physical activity, higher BMI, and smoking increase the mortality risk of breast cancer patients.

Screening

Recommendations regarding screening vary from country to country. The targeted age, the interval between two screenings, the method to define the risk and even the real absolute benefit of screening programs are still debated. It seems, however, that screen-detected breast cancer is associated with a better prognosis compared to cancers diagnosed through symptomatically [46], even after accounting for lead-time and length-time biases [48–50] . This can be explained by smaller tumour size and less nodal involvement in screen-detected women.

Tumour characteristics

Tumour stage

Tumour stage is defined by the TNM classification [51]. It combines the size of the tumour, the nodal involvement and the presence of metastases. Stage I (localised tumour) has a much better prognosis than a stage IV tumour (distant disease).

Tumour size

Tumour size (T) is defined as the largest diameter of the primary breast tumour. Larger tumours have worse prognosis. Tumour size is one of the strongest predictors of survival, even a long time after diagnosis. Tumour size is often correlated with nodal involvement. We distinguish T0 (no evidence of primary

tumour), Tis (Tumour in situ), T1 (tumour is 2cm or less in greatest dimension), T2 (tumour is between 2 and 5cm in greatest dimension), T3 (tumour is more than 5cm in greatest dimension) and T4 (Tumour of any size with extension to chest wall and/or skin).

Regional lymph node involvement

Lymph node involvement describes the number of ipsilateral nodes with metastatic tumour growth. It is a strong prognostic factor, with node positive patients having on average 4 to 8 times higher mortality compared to those who do not have any nodes affected [51]. At a pathological level we distinguish N0 (No regional lymph node metastasis), N1 (Micrometastases or metastases in 1-3 axillary lymph nodes and/or in internal mammary nodes, with metastases detected by sentinel lymph node biopsy but not clinically detected), N2 (Metastases in 4-9 axillary lymph nodes, or in clinically detected internal mammary lymph nodes in the absence of axillary lymph node metastases) and N3 (Metastases in ≥ 10 axillary lymph nodes or in infraclavicular lymph nodes, or in clinically detected ipsilateral internal mammary lymph nodes in the presence of ≥ 1 positive level I, II axillary lymph nodes, or in > 3 axillary lymph nodes and in internal mammary lymph nodes, with micrometastases or macrometastases detected by sentinel lymph node biopsy but not clinically detected†, or in ipsilateral supraclavicular lymph nodes).

Metastasis

Patients with metastasis present poorer survival compared to those who have not. For example, a study predicted 20-year survival of patients with a local breast cancer (node-negative, tumour size < 1 cm, aged >=45 and grade 1) to

be 89%, whereas for patients with metastasis it was 0.18% [52]. We distinguish M0 (no distant metastasis) from M1 (presence of distant metastasis).

Histological subtype (morphology)

The six most common types of breast cancer are as infiltrating ductal carcinoma, infiltrating lobular carcinoma, mucinous or colloid breast carcinoma, medullary breast carcinoma, tubular breast carcinoma and inflammatory breast carcinoma. The disease prognosis varies substantially according to the morphology of the tumour. Tubular, mucinous and medullary breast carcinomas have a better prognosis than the other sub-types. Inflammatory breast cancer has a very poor prognosis (10-year survival <30%) [53].

Histological grade

The grade of a breast cancer tumour is assigned using the Elston-Ellis grading system. It represents the degree of tumour differentiation, that is, the degree to which the tumour no longer looks like the original tissue. It is based on a combination of architectural and nuclear characteristics. More precisely, tubule formation, nuclear pleomorphism and mitotic activity are each scored on a scale of 1 to 3.. The sum of these scores represent the overall grade [32]. Higher grades (poorly differentiated tumours: total score 8 or 9) have been consistently associated with lower long-term survival [54]. Grade 2 (moderately differentiated tumours: total score 6 or 7) offers a lack of prognostic information as it probably consists of a mix of biologically low and high-grade tumours. For this reason, the Breast Task Force of The American Joint Committee on Cancer has chosen not to include grade in the revised TNM staging system for breast cancer.

Hormone receptors

Breast cancer cells either have, or not have, receptors for the hormones oestrogen and progesterone. Hormone receptors are proteins, found in and on breast cells, whose role is to pick up hormone signals telling the cells to grow. A cancer is called oestrogen-receptor-positive (ER+) if it has receptors for oestrogen. This suggests that the cancer cells, like normal breast cells, may receive signals from oestrogen that could promote their growth. Similarly, the cancer is progesterone-receptor-positive (PR+) if it has progesterone receptors. Again, this means that the cancer cells may receive signals from progesterone that could promote their growth.

According to the presence of receptors, the cancer is likely to respond to hormonal therapy or other treatments. Hormonal therapy includes medications that either lower the amount of oestrogen in the body or block oestrogen from supporting the growth and function of breast cells. If the breast cancer cells have hormone receptors, then hormonal medications can help to slow or even stop their growth. If the cancer is hormone-receptor-negative (no receptor is present), then hormonal therapy is unlikely to work. Thus, breast cancer survival is positively associated with the ER and PR levels [55].

HER-2 Expression

HER-2 is defined as the Human Epidermal growth factor-2 Receptor. Patients whose breast cancer cells present amplification and/or an overexpression of HER-2 have a poorer prognosis compared to patients who do not have this amplification/over-expression. (10-year survival 50 vs. 65%) [55].

Bioscore

In addition to their staging system, the American Joint Committee on cancer developed a score (Bioscore) to measure patient's stage. Briefly, this system considers grade, ER status, and HER2 status. Grade 1 or 2 tumors, ER-positive tumors, or HER2-positive tumors are assigned a score of 0. Grade 3 tumors, ER-negative tumors, and HER2-negative tumors are scored with one point. Added together, a risk-profile Bioscore from 0 to 3 can then be calculated. Within each TNM stage, the risk Bioscore can be used to further stratify patients [56].

Other prognostic factors

There are a lot of other molecular factors which are prognostic and will be of great importance in the future. These include a woman's genomic profile, gene expression profile, or markers of proliferation of the tumour cells. However, they are largely unavailable for population-based data.

All these factors are essential to understand the mortality patterns of long-term survivors and some of them will be used in my research towards this goal.

Treatment

Depending on the characteristics of the tumour a large range of treatment is available for breast cancer. Four different therapies, used alone or in combination might be suitable: surgery, radiotherapy, chemotherapy, and hormonal therapy.

Surgery consists of mastectomy or lumpectomy (also called breast conserving surgery). It offers a local treatment to remove the primary tumour. Often, lumpectomy needs to be combined with post-operative radiotherapy. Radiotherapy uses high-energy radiation, externally or internally, to irradiate remaining cancerous cells. Both breast and lymphatic tissues can be treated.

Chemotherapy is often used too. Chemotherapy is the use of systemic drugs to combat the disease and stop cancer cells from growing. This might be administrated either before the surgery, as a neo-adjuvant therapy, to reduce the tumour size, or after the main treatment, as an adjuvant therapy. Radiotherapy and chemotherapy are non-discriminatory treatments affecting both normal and abnormal cells and lead to the death of healthy tissue as well as the cancer. Hormonal treatment is a more targeted treatment. Its effectiveness depends on the hormone receptor status of the patients' tumour. Hormone therapy inhibits hormone production, which inhibits the growth and proliferation of the tumour in the presence of hormone receptors in the tumour.

BOX 2

Summary

- ✓ Breast cancer is a complex and multifactorial disease.
- ✓ Multiple factors have an impact on the prognosis of the disease.

What is next?

- ✓ Breast cancer therefore needs a particular type of surveillance.

Surveillance of breast cancer

Population-based data

Effective cancer control draws on results of a combination of clinical trials and observational studies. The two are often placed in opposition with each other but should in fact be considered as complementary [57].

Clinical trials are always subject to randomization. This process induces good internal validity, meaning that the observed phenomenon is real and very unlikely to be due to chance, bias or confounding. This enables the effect of a particular intervention or drug to be fairly assessed. However, clinical trials inform us about efficacy in the research setting and do not necessarily describe the overall management of cancer patients in the general population setting. Indeed, less than 10% of patients with cancer are enrolled in a trial, and generally these patients are very different from patients treated in routine practice. Trials tend to include only very specifically selected individuals (patients with advanced age, comorbidities are usually excluded) and are, therefore, not representative of the whole population of cancer patients.

In particular, observational studies provide information about cancer control in a whole population and allow investigators and policy makers to evaluate the effectiveness of the health system provided for the general population, as well as providing insight into both short and long-term toxicity of cancer treatments. Observational studies include all patients within a given jurisdiction and are therefore less prone to selection and referral biases. Large-scale observational studies have been enabled by advances in computer technology that provide linkages between separately collected routine databases.

Observational studies do, nevertheless, have important limitations that must be carefully considered when evaluating any effect. We distinguish selection bias, information bias and the most challenging, confounding. Selection bias happens when the study population does not represent the target population. Information bias is related to incorrect measures of exposure and/or outcome of the study. Confounding happens when the relationship found between the outcome and the exposure can be explained by the distribution of a third variable, called confounding factor. All these biases need to be discussed when considering the results of observational studies.

Cancer registries

Population-based cancer registries collect population-based data and so monitor progress against cancer. The first was created in Hamburg (Germany) in 1926. Today at least 290 population-based cancer registries exist worldwide. Their role is the systematic collection, storage, analysis, interpretation and reporting of cancer data on patients with cancer [58]. These data are essential to evaluate the current situation, to set objectives, to define priorities, and to assess the future evolution of the disease burden [60,61].

Cancer registries play an important role in this research, both by providing routine data on patterns and trends, as well as data for analytic epidemiological studies. More recently, cancer registries have progressively developed their activities to include data on time trends and geographical variations and survival from cancer. This information allows monitoring of progress in implementing cancer control activities, and evaluation of prevention, early detection, screening and treatment interventions.

For each new (i.e. incident) cancer case, registries record at least some details of the individual affected and the nature of the cancer. Cancer registrars, who are data information specialists skilled at capturing a complete history for every cancer patient, perform the collection. The vital status of the patient can be updated regularly via linkage with routine death registrations in order to maintain accurate surveillance information. Lifetime follow-up data from patients permit registries to analyse cancer patient survival of the population through time.

The role of a population-based cancer registry is thus pivotal. Epidemiological data obtained from cancer registration are indispensable to document population-based trends in patient numbers; ascertain the need for prevention, screening and therapeutic measures; allocate targeted resources for research; document the quality and efficiency of cancer treatments; and conduct causal research.

Data quality

In view of the large responsibilities that cancer registries hold, it is crucial the data they collect are of high quality. Even if the data are analysed with the best and most up-to-date methodologies, results and conclusions can be severely compromised by a lack of data quality.

For cancer registries, we distinguish four dimensions in terms of data quality:

- ✓ Comparability

  This allows comparisons to be made between populations and over time. To achieve this, standardization of practices is required regarding coding, classification and definitions. The World Health Organization

(WHO) proposed the "International Codification of Disease - Oncology (ICD-O)" to provide the international standards.

- ✓ Timeliness

  This is related to the rapidity of data reporting. The aim is minimizing the delay between the occurrence of a cancer and its recording.

- ✓ Completeness

  This is defined by the extent to which all eligible cases have been registered. A measure for the evaluation of completeness is the proportion of cases that have been recorded through their death certificates.

- ✓ Validity:

  This concerns the accuracy of the recorded data.

BOX 3

Summary

- ✓ Population-based data about breast cancer are collected worldwide through cancer registries
- ✓ These data allow surveillance of the disease

What is next?

- ✓ Accurate follow-up data allow the study of survival, a key indicator

Survival, a key indicator

Definition

There is a great interest, from policy makers, clinicians treating cancer, as well as from patients, in the overall evaluation of progress against cancer. This is possible through cancer registries and the use of their data [60]. Together with incidence and mortality, survival is a key indicator used for this purpose at a population-based level [61]. Despite potential biases such as lead-time or length time, survival of a particular group of patients diagnosed in the same period with the same disease provides essential prognostic information for newly diagnosed patients.

The main aim of survival analyses is to describe the proportion of study participants who have experienced a specific event of interest at a particular point in time after a set starting point. The data involved in survival analyses, failure-time data, are characterized by a starting time (diagnosis of cancer), and an end time defined by the occurrence of the event of interest (death from cancer, recurrences or other event related to the disease) or the end of follow-up. This type of analyses is unique because of the presence of censoring. In the context of cancer epidemiology, survival is estimated for a group of cancer patients by following them from their diagnosis until death (the event of interest) or until the end of the study [61–63].

Several reasons could lead to the outcome not being observed during the study's follow-up period (censoring). Individuals could be lost to follow-up, or a competitive event (death from another cause) could happen earlier than the

outcome under interest (death from cancer). In order to deal with this type of data, specific methodologies are needed.

Net survival

It is crucial, when planning to analyse the survival experience of cancer patients, to decide which particular quantity we would like to estimate with my data [65].

The simplest measure of survival is overall survival, when all deaths, whatever the cause, are considered as events. It is defined by the probability that a patient is still alive at a certain time point after the diagnosis. Although straightforward, this concept has a limited interest insofar as it does not give any specific information about a cause-specific prognosis.

The alternative concept is the survival related to the disease of interest, disregarding deaths from other causes [65]. Here the outcome of interest is death due to breast cancer (direct or indirect deaths). Deaths due to other causes are therefore considered as competing events. This concept splits the overall mortality into the deaths that can be attributed to the disease of interest and those due to other causes.

As an illustration of the differences between overall and disease-specific survival related to the disease, we could make a comparison for patients diagnosed with breast cancer divided in two age groups. The results are summarised in Table 1.

	Overall survival (%)	Disease-specific survival (%)
Young age	88	88
Older age	33	88

Table 1: 5-year overall and disease-specific survival by age groups for patients diagnosed with breast cancer.

The older patients have clearly a lower overall survival. However, the disease-specific measure shows that the mortality related to the disease is similar for both age groups.

These two types of survival lead to the consideration of two separate worlds: the real world and the virtual or hypothetical world. In the real world, we take into consideration deaths from other causes as they occur. The probability of death due to breast cancer is therefore the probability that a patient dies as a consequence of their disease before another cause. In contrast, in the hypothetical world, we eliminate deaths from other causes. Net survival is the probability of survival from breast cancer in a situation where dying from causes other than the cancer under study is not possible [62].

'Real world' estimations of survival are of interest in the exchange between clinicians and patients, and for medical decision making, as these statistics provide information on the overall risk of a particular patient dying from the cancer during a specified period of time, in a particular locality.

However, when comparisons are required between populations, including between different age groups, net survival and the 'hypothetical world' is the better approach because it is the only measure that is independent of the differences in the mortality from other causes. As such, net survival allows us to evaluate research questions related to disease aetiology: the results reflect differences in survival associated only with the exposure under study as opposed to a mixture of disease-specific mortality and non-cancer mortality [66].

Two data settings

Two settings have been used for the estimation of net (disease-specific) survival.

The cause-specific setting relies on information about the underlying cause of death for each patient. Deaths due to causes other than the disease of interest (here, breast cancer) are considered censored in the survival analyses. The survival for a given time period can then be calculated directly using standard methods such as Kaplan-Meier [64,67].

The alternative is the relative-survival setting. Relative survival is defined as the ratio of the observed survival rate in a specified group of patients, during a specified period of time, to the expected survival rate, which is the expected rate of an exactly comparable group of individual in the population from which the cancer patients are drawn [68]. Ideally, the disease under study should be excluded from the general population mortality rates estimation, although in practice this will make little difference 68.

Unlike the cause-specific setting, the relative survival setting does not require reliable information about cause of death. In its place, estimated rates of background, or expected mortality are used to off-set the overall mortality and provide an estimate of the excess mortality due to the cancer. This estimate disregards information on whether the death is directly or indirectly related to breast cancer.

Net survival within the relative survival setting is based on the idea that the mortality can be divided in two additive processes, which are independent. The first is mortality linked to deaths from the disease under study, breast cancer, and the second is the mortality related to deaths from all other causes. Considering instantaneous hazard rates, we have,

$$\lambda_o(t) = \lambda_p(t) + \lambda_e(t)$$

where $\lambda_o$represents the overall mortality in the cohort, λ_p defines the mortality linked to the other causes of death or the population mortality and λ_e is the mortality linked to breast cancer only (also called the excess mortality). If we integrate over t, we have,

$$\int_0^t \lambda_o(u)du = \int_0^t [\lambda_p(u) + \lambda_e(u)]du$$

Thus,

$$exp\left(-\int_0^t \lambda_o(u)du\right) = exp\left(-\int_0^t \lambda_p(u)du\right) exp\left(-\int_o^t \lambda_e(u)du\right)$$

And finally, using the mathematical definition of survival, we find

$$S_o(t) = S_p(t).S_e(t)$$

Or,

$S_e(t) = \frac{S_o(t)}{S_p(t)}$Survival probabilities in the cancer group $S_o(t)$ and external group $S_p(t)$ are therefore compared along time since diagnosis. A lower survival in the cancer group represents higher cancer-related mortality in that group.

BOX 4

Summary

- ✓ Net survival is a key indicator for the surveillance and understanding of progress against breast cancer
- ✓ Two data settings have been used for its estimation

What is next?

- ✓ The Geneva Cancer Registry is a unique context in which net survival can be evaluated

The context of Geneva

Burden of disease

The incidence rate for breast cancer in the canton of Geneva is one of the highest in the world. This is likely to be due to the fact that Geneva is particularly representative of a western lifestyle. Trends showed an increase in incidence in the last two decades, partly associated with the introduction of screening programs. More particularly, it induced an increase in the 50-70-year age group of more localized tumours. Lately, because of a decreasing use of hormone replacement therapy, the incidence rate appears to be decreasing (Figure 6).

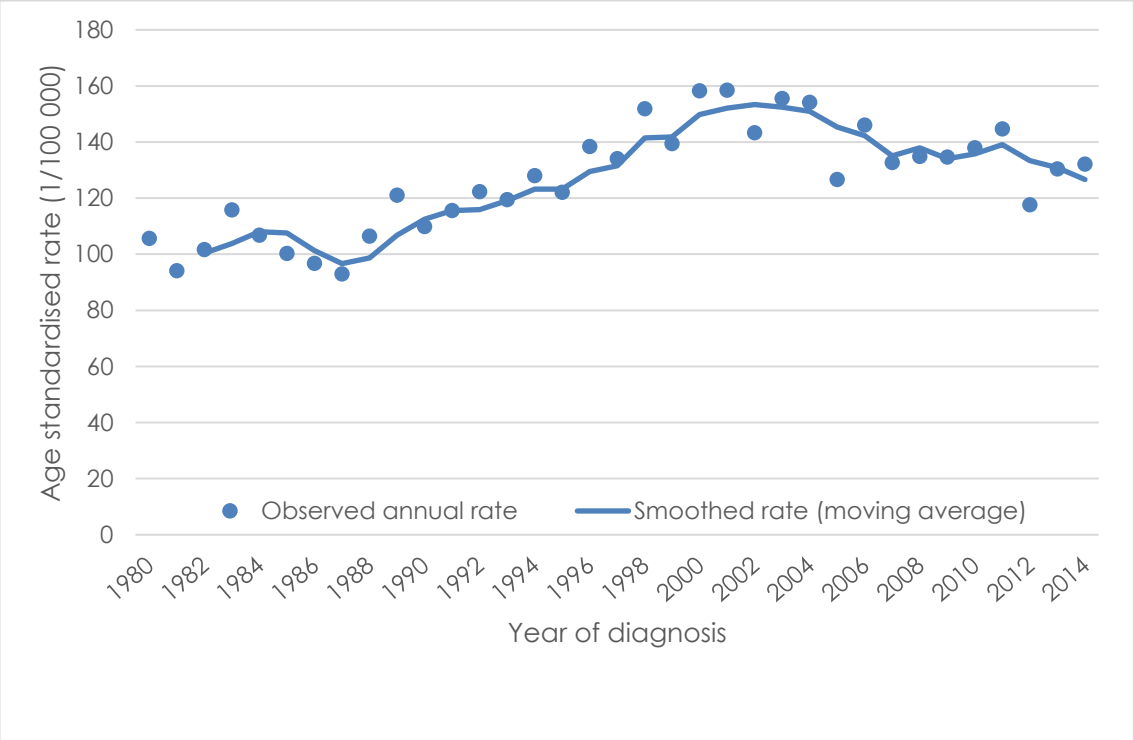


Figure 6: Age standardised incidence rate of breast cancer in the canton of Geneva by period of diagnosis, females, 1989-2013. Source: Geneva Cancer Registry

Breast cancer survival is amongst the highest of all cancer sites and has dramatically increased with time. Indeed, in Geneva the 5-year disease-specific survival rate ranged from 45% in the 1970s up to 84% for patients

diagnosed between 2000 and 2009 (Figure 7). The prevalence, which represents the number of women living in Geneva who have ever had a breast cancer, i.e. survivors of the disease, is presented for Geneva in Figure 8 . It also increases each year, from 1.1% in 1990 to 2.4% in 2014.

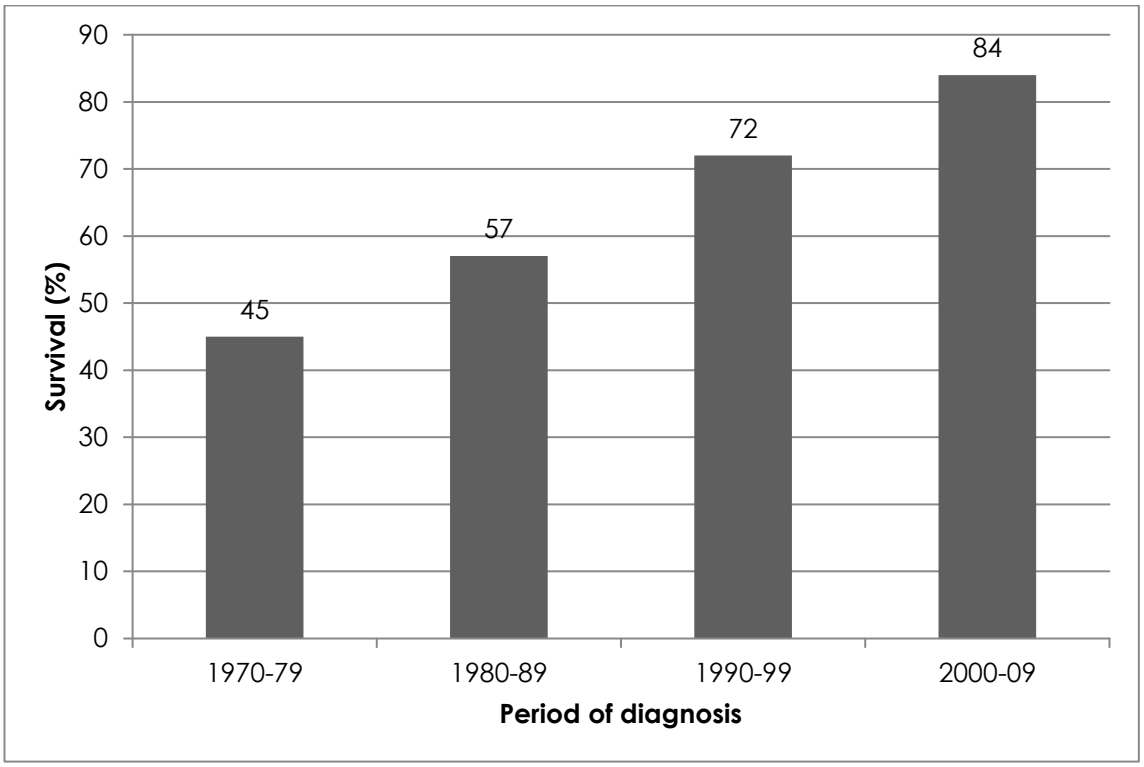


Figure 7: Disease-specific survival trends for cancer patients diagnosed with breast cancer in Geneva. Source: Geneva Cancer Registry.

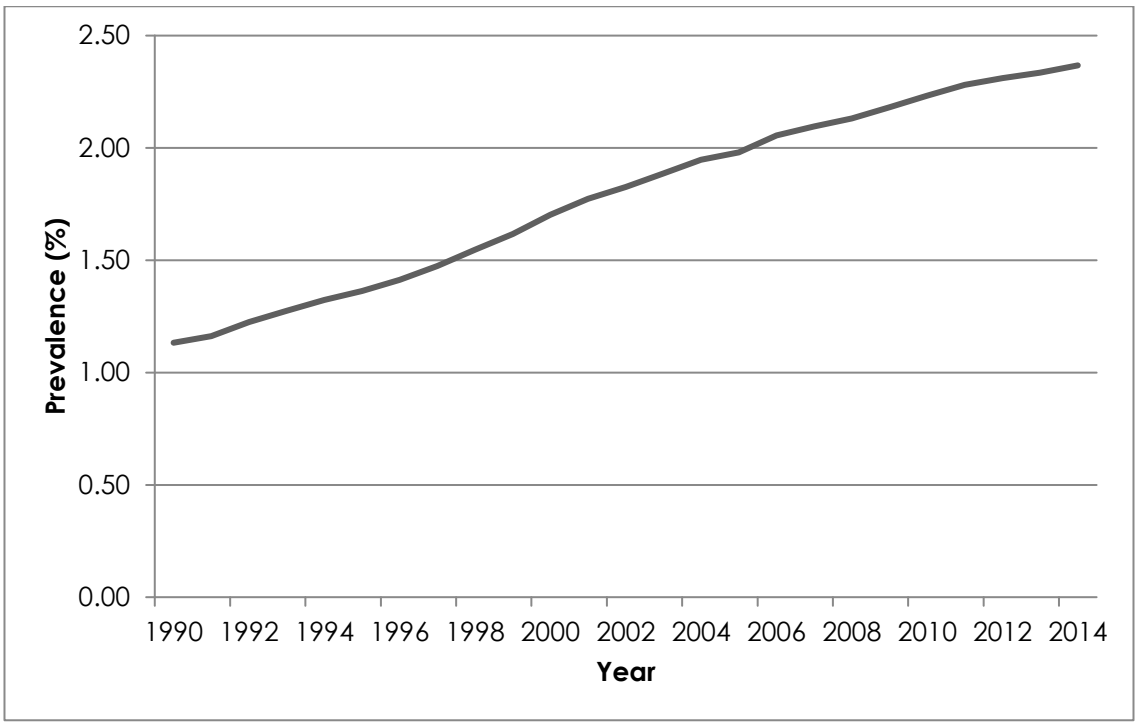


Figure 8: Trend of breast cancer prevalence in Geneva, 1990-2014. Source: Geneva Cancer Registry.

Geneva Cancer Registry

The Geneva Cancer Registry, the first registry in Switzerland, was created in 1970 following a decision of the Council of State. One of the first periodical reports of the Registry stated its tasks:

- ✓ Collect information on all neoplastic diseases detected among the resident population in the canton, in order to obtain epidemiological statistics (calculation of the incidence and prevalence, calculation of survival, etc.), establish the level of impact of these conditions and measure the stage of the lesion at the time of its discovery in the general population or in specific subgroups. The identification of high-risk groups or those for whom an early diagnosis is relatively infrequent serves both the preventive and the etiological research functions.

- ✓ Follow up the cases reported for the evaluation of survival from major tumour types and for calculating the intervals between significant events in the history of the lesion (overall evaluation of preventive and therapeutic measures).

One of the early objectives of the registry also included the organization of a system for the regular exchange of information with the physicians responsible for monitoring cancer patients. During the first two years of registration 1,112 cases of cancers among males and 1,227 among females were recorded in the Canton (in 1970, Geneva had 331,599 inhabitants in an area of 282 km^2). More than forty years later, at the end of 2016, these figures had become 66,600 and 71,200 for males and females, respectively (by 2016, the number of inhabitants in the Canton was 470,512).

The registry has continuously collected exhaustive information about patients diagnosed, treated or dying from a cancer among the Geneva canton population. All cancers arising in resident patients are recorded even if the cancer is treated outside the canton. During these 40 years, a lot of changes and improvements have been made in the organization and management of the registry as well in its functions. Not only had the number of cases that were registered increased substantially but the number of variables recorded for each case went from 38 to the current 150.

Up to 2016, more than 100,000 tumours had been recorded in the Registry's database, along with variables concerning individual and tumour characteristics and treatment, collected from various sources. The Cancer Registry has a very low percentage (<2%) of cases recorded from death certificates only, indicating that it is very complete. All hospitals, pathology

laboratories and private practitioners in the canton are requested to report all cancer cases. Trained tumour registrars systematically abstract data from medical and laboratory records. Physicians regularly receive enquiry forms to complete missing clinical and therapeutic data.

Follow-up of all patients is provided by the OCP (*Office Cantonal de la Population)*. The OCP, the office in charge of the registration of the resident population (approximately 450,000 inhabitants, mainly urban) performs a close active follow-up of the whole population of Geneva. Arrivals and departures in Geneva for work, stay or living, have to be declared. All death certificates, after being filled by general practitioners, are centralized in Neuchatel (at the *Office Federal de la Statistique*, the Federal Office for Statistics). All these data are shared annually with the Geneva Cancer Registry and allows an active and exhaustive follow-up of incident cases.

Dual coding of cause of death

Uniquely amongst cancer registries, the Geneva Cancer Registry records two different variables pertaining to cause of death. The first is based on the information available on the death certificates and the second on a comprehensive clerical review of the patient's medical records using all available clinical information available. This dual recording of death provides the unique context in which my study is able to evaluate the best means of estimating long-term net survival.

A unique context

Geneva is representative of the burden that is observed in developed countries. Women are living longer than ever after being diagnosed with breast cancer, because of better diagnostic tests and improvement in treatment

strategies. As a result, the number of breast cancer survivors is increasing, leading to a much greater interest in long-term survival. To be able to measure this long-term survival, the Geneva Cancer Registry offers more than 40 years of follow-up. Moreover, thanks to the clerical review of the cause of death, it will be possible to conduct an accurate comparison between the two data settings available for this estimation of net survival. Finally, because of the numerous and high-quality data that are recorded, it is possible to consider the evaluation of a significant number of prognostic factors on long-term net survival.

Summary box

- ✓ The burden of breast cancer is very important worldwide.
- ✓ Breast cancer is, in developed countries, the most common malignancy and cause of cancer death amongst women.
- ✓ This hormone-dependent tumour presents, however, good prognosis, thanks to improvement in its monitoring and surveillance.
- ✓ Population-based cancer registries take part in surveillance by providing observational data about cancer, in particular, data regarding follow-up which are used for the estimation of survival.
- ✓ Because the number of breast cancer survivors is increasing, there is a growing interest in gaining a better understanding of long-term survival.
- ✓ Net survival is the concept that allows accurate evaluation from the disease of interest.
- ✓ Net survival can be derived using either the cause-specific setting, for which the underlying cause of death is required, or the relative-survival setting, which compares the overall survival of the cohort of patient to that they would have experienced if they had had the same mortality experience of the general population.
- ✓ The Geneva Cancer Registry represents a suitable place for this study because of high-quality data with long follow-up.

[CHAPTER ONE]

Data quality

Net survival can be derived using both the cause-specific and relative-survival setting. In this first chapter I focus solely on the cause-specific setting and in particular the question of data quality.

Using the cause-specific setting implies the availability of information on the underlying cause of death for all deceased cancer patients.

This Chapter completes the first aim of the thesis, which is the evaluation of the accuracy of the cause of death information routinely collected within a cancer registry setting and the impact of inaccuracies upon survival estimates.

In order to achieve this aim, I define the following objectives:

- to investigate how accurate, the routinely recorded cause of death field is compared to the cause of death field derived from comprehensive clerical review and

- to compare cause-specific survival estimates of net survival using routinely recorded cause of death to those derived using clerically reviewed cause of death.

This chapter comprises a copy of the research paper published in *BMC Cancer* alongside text, which describes the background to these objectives, summarises the approach and findings, and specifies how the paper fulfils the aim and objectives.

Background

Survival within the cause-specific setting

When estimating survival within the cause-specific setting only deaths due to the disease of interest, breast cancer in our case, are counted as events. Patients dying due to other underlying causes are censored at their time of death. In order to estimate net survival in the cause-specific setting, information on the underlying cause of death for every patient is thus required. Statistical tools available for overall survival, such as Kaplan-Meier [63] and the semi-parametric Cox model [70] can then be applied to the censored data in order to estimate survival related to the disease.

Accuracy of cause of death

The accuracy of the underlying cause of death itself is thus extremely important in order to obtain a valid estimate of net survival within the cause-specific setting. However, this accuracy cannot be assumed. Berkson and Gage said that "The determination of whether a death is entirely due to cancer or entirely due to other causes is difficult to establish, if indeed it is even possible to define precisely. Actually in most cases it is impossible to establish unequivocally…" [63]. In cause-specific settings, the difference between a death directly caused by cancer and a death being an indirect consequence of it is difficult to determine and could lead to misclassification and erroneous survival estimates. Percy *et al.* [71] were the first to demonstrate the impact of misclassification on mortality statistics and showed that only 65% of death certificates were reporting the cause of death that was described in hospital notes.

The Geneva Cancer Registry

The Geneva Cancer Registry has the particularity not just to record routinely information from death certificates but also to perform a clerical review of the underlying cause of death for each patient using all the clinical information available in their medical records. This information is gathered from the death certificates themselves, autopsy reports, the 'letter at death' written by the patient's general practitioner and all the patient's medical notes. By this process a revised cause of death variable is obtained. This dual recording of cause of death in Geneva provides an ideal setting within which to examine how accurate death certificates actually are, and the extent of the potential bias introduced to survival estimates by using routinely recorded deaths for the measurement of underlying cause.

Paper One

The reviewed cause of death

Description

"**Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva Cancer Registry**" was published in *BMC Cancer* in 2013 [71]. In it, I describe the process of cancer registration in Geneva. I compare causes of death derived from death certificates with the cause of death derived from clerical review and calculate the Kappa statistic to estimate the degree of correspondence between the two different variables. I also describe the most common errors between the two variables, taking the validated version as the 'gold standard'. I finally conduct an analysis, which compares the survival related to the disease that I would observe 1) when using routinely recorded cause of death and 2) when clinically reviewed cause of death was used.

Main results

I report that both recording processes matched perfectly for 95.8% of the cohort. 2.5% of the cases were considered to have died from breast cancer according to their death certificates and but were recoded with a different underlying cause of death by the registry. Among these women, the cause of death was mostly recoded to heart disease (48%) and other malignant tumours (20%). On the other hand, 1.7% of the cases were recorded as dying from a cause other than breast cancer according to the death certificate but were recoded to death from breast cancer by the registry. Among these women, the main causes of death reported on their original death certificates were other malignant tumours (40%) or an imprecise code (19%). The value of the kappa

test reached 0.82 and was statistically significant. Unadjusted concordance varied greatly between subgroups. The concordance was significantly lower with increasing age. Similar age-related trends, though not significant, were found among the three subpopulations defined by time since diagnosis. These age-related patterns were much less marked for age at death. Concordance was greater for early stage of disease (stage I and II) compared to advanced stage (III and IV). A clear pattern was found according to the type of treatment with higher concordance for complete, with curative intent, treatment (0.83), intermediate concordance for palliative treatment and lower concordance for non-treated patients. The odds of disagreement increased significantly with age at diagnosis. We observed the same trend when using age at death. Period of diagnosis was not significantly associated with disagreement but we did observe a significant decreasing trend for period of death as a continuous variable. Patients treated palliatively had significantly higher odds of disagreement. Patients treated in the public sector also had a higher risk of disagreement as well as those who died in a public hospital. The survival curves matched almost perfectly, with a difference in 20-year survival lower than 1%. We observed substantial differences for several subgroups: patients aged 70–79 or 80+, patients diagnosed in the two first period of diagnosis, patients with no treatment, patients with hormonal therapy, and patients with metastatic tumours.

Conclusion

The overall concordance between the two types of death recording was high, and the impact on short- and medium-term survival for the whole cohort fairly minimal. This suggests that a substantial part of the information related to death is captured by the routinely recorded cause of death. Nevertheless, for some

subgroups, the additional information use to better define the underlying cause of death was shown to be useful because survival estimates were more divergent. This was particularly true for older patients and patients presenting with more advanced disease. The concordance between the two variables was also observed to decrease with time since diagnosis. Thus, the availability of additional clinical information appears to be progressively more important for the determination of the underlying cause of death as follow-up time increases, suggesting that reviewed cause of death is more accurate for the estimation of long-term survival.

Fulfilment of Aims and Objectives

In this Chapter I have evaluated how accurate the recording of the cause of death at the Geneva Cancer Registry was in this period, and the impact of inaccuracies upon survival estimates.

Like all cancer registries worldwide, the Geneva Cancer Registry uses the International Classification of Disease (ICD-10) [72] to record the underlying cause of death. Unlike other registries, it also provides a revised cause of death using all available clinical information.

I have shown that there is an overall concordance between these two variables, but that important differences are present and these have an impact upon survival estimates. As a result of these analyses I conclude that the clerically reviewed cause of death is more accurate for the estimation of cause-specific survival. Even if the advantage is limited for the whole cohort in the short- and medium-term, revised cause of death provides much more accurate estimation of survival for sub-populations and, importantly, of long-term survival for all patients, and should thus be used in cause-specific analyses.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Robin Schaffar
Principal Supervisor	Dr Laura Woods
Thesis Title	Long-term net survival among women diagnosed with breast cancer. Accuracy of its estimation and evaluation of its determinants

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	BMC Cancer		
When was the work published?	2013		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. I planned and carried out the literature review and data analysis. I prepared all drafts of the paper. The co-authors provided input and feedback on the content data analysis and on the paper drafts prepared by me.

Student Signature: [redacted] **Date:** 09/03/2018

Supervisor Signature: _____ **Date:** 09/03/2018

BMC Cancer

RESEARCH ARTICLE Open Access

Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva cancer registry

Robin Schaffar[1,2*], Elisabetta Rapiti[1], Bernard Rachet[2,3] and Laura Woods[2]

Abstract

Background: Information on the underlying cause of death of cancer patients is of interest because it can be used to estimate net survival. The population-based Geneva Cancer Registry is unique because registrars are able to review the official cause of death. This study aims to describe the difference between the official and revised cause-of-death variables and the impact on cancer survival estimates.

Methods: The recording process for each cause of death variable is summarised. We describe the differences between the two cause-of-death variables for the 5,065 deceased patients out of the 10,534 women diagnosed with breast cancer between 1970 and 2009. The Kappa statistic and logistic regression are applied to evaluate the degree of concordance. The impact of discordance on cause-specific survival is examined using the Kaplan Meier method.

Results: The overall agreement between the two variables was high. However, several subgroups presented a lower concordance, suggesting differences in calendar time and less attention given to older patients and more advanced diseases. Similarly, the impact of discordance on cause-specific survival was small on overall survival but larger for several subgroups.

Conclusion: Estimation of cancer-specific survival could therefore be prone to bias when using the official cause of death. Breast cancer is not the more lethal cancer and our results can certainly not be generalised to more lethal tumours.

Keywords: Cause-specific survival, Cause-of-death, Cancer registry, Concordance

Background

Population-based cancer survival is widely used to evaluate the impact of health care systems in disease management. Net survival is the survival that would be observed if the only possible cause of death were the cancer of interest [1]. Net survival is especially relevant when the cohort of interest become older since the risk of dying from other causes than cancer increases. Net survival is also very useful when comparing subgroups whose mortality due to other causes could be different and therefore lead to biased estimation of the survival contrast.

Two main data designs can be distinguished, the cause-specific and the relative survival designs, according to the availability of information on cause of death. Such information is rarely available in routine, population-based data and net survival is then commonly estimated within the relative survival framework. However, when information about the underlying cause of death is available, net survival can be estimated using the cause-specific approach, in which only deaths from the cause of interest are considered as 'failures', while deaths from other causes are

* Correspondence: robin.schaffar@lshtm.ac.uk
[1]Geneva Cancer Registry, Institute for Social and Preventive Medicine, University of Geneva, 55 Boulevard de la Cluse, Geneva, 1205, Switzerland
[2]Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK
Full list of author information is available at the end of the article

censored. High-quality information on the cause of death is required for each individual patient. This information is commonly available only in clinical trials or hospital series, but the cause-specific approach is sometimes used on population-based data from cancer registries, where the underlying cause of death is derived from death certificates. The underlying cause of death is the "disease or injury which initiated the train of morbid events leading directly to death" or the "circumstances of the accident or violence which produced the fatal injury". It is codified in The International Classification of Diseases (ICD), which was designed to classify causes of death for statistical tabulation and research. Despite these international rules (developed over 100 years), comparability and accuracy issues still arise. Different medical terminologies, inaccurate completion of the death certificates, misinterpretation or misapplication of the coding rules for selection of the underlying cause of death can cause comparability problems between different geographical areas and/or different periods of time. The validity and accuracy of the reported underlying cause of death may also be incorrect if the clinician's certification does not accurately reflect the clinical history of events leading to death.

Percy et al. were the first to report that misclassification of the underlying cause of death could bias the mortality trends and therefore the estimation of cancer-specific survival [2]. Many other studies, then, have highlighted the issue of inaccuracy of the cause of death information obtained from death certificates [3-9]. Some studies have shown that the proportion of misclassification can be very high [4,10]. However, one study has suggested that the proportion of misclassification can be lower for screened patients dying from breast cancer [11].

The validity of disease-specific survival is based on the assumption that the underlying cause of death is accurately determined. The Geneva Cancer Registry, which collects all the death certificates of routinely recorded deaths in the Geneva canton (Switzerland), also reviews the cause of death of each registered cancer patient using all the available clinical information relating to the patient's disease and treatment. This leads to a particular and unique situation in which a second, validated variable defining the cause of death is generated. This second variable is considered to be a more reliable record of the patient's cause of death and so will be expected to give rise to more accurate estimates of cause-specific survival.

The purposes of this study are (a) to describe the process of recording the cause of death in the Geneva Cancer Registry, (b) to investigate how accurate the routinely recorded cause of death is compared to the validated cause of death derived from clerical review and (c) to examine whether the process of validation leads to differences in the estimates of cause-specific survival.

Methods

Data

The data used in this study were obtained from the Geneva Cancer Registry. All women diagnosed with a breast cancer between 1970 and 2009 and resident in Geneva were included in the study.

The Geneva Cancer Registry collects information on incident cancer cases from various sources, including hospitals, laboratories and private clinics, all requested to report new cancer cases. Trained registrars systematically extract information from the medical records and conduct further investigations in the case of missing key data. The variables of interest for this study were cause of death as specified on the death certificate, revised cause of death, age at diagnosis, age at death, year of diagnosis, year of death, social class, stage of the tumour, treatment, sector of care and place of death. The Geneva Cancer Registry has general registry approval by the Swiss Federal Commission of Experts for professional secrecy in medical research (Commission d'experts pour le secret professionnel en matière de recherche medical). This approval permits cancer data collection and its use for research purposes.

Coding of cause of death

The Geneva Cancer Registry is notified of all deaths occurring in the Geneva canton through three different processes.

First, when a patient dies in the canton of Geneva, a death certificate is compulsorily completed by the clinician certifying the death who reports the primary, secondary and concomitant causes of death. The Geneva Cancer Registry receives photocopies of all these death certificates through the Geneva Health Administration; and links them to the incidence database. The causes of death reported on the death certificates represent the original causes of death.

Meanwhile, once a year, the Federal Office of Statistics (Office Federal de la Statistique, OFS) which is a national publicly-funded organisation collecting death certificates and maintaining a mortality database for the whole of Switzerland provides the Geneva Cancer Registry with a mortality database for the Geneva canton. This is also linked to the incidence database to complete and/or validate the process described above. This leads to the definition of the official cause of death as the underlying cause of death derived from death certificates.

Finally, the Geneva Cancer Registry is provided on an annual basis with information on the vital status of the Canton population by the Cantonal Office of the Population (Office Cantonal de la Population, OCP). OCP is a regional administration that monitors births, deaths, migration, residency and civil partnerships. Only information about the vital status of a patient (deceased or not),

or information on whether a person has migrated from Geneva is provided to the Registry. Information on the cause of death is not available within this database.

After all the records are merged, the Cancer Registry registrars then go back to the patient's charts and review the cause of death according to all the documents available. These include death certificates, autopsy reports, letter at death written by general practitioners and all the patient's medical notes. By this process the cause of death variable, the *revised cause of death,* is obtained.

Sometimes, the Geneva Cancer Registry is able to obtain information about the occurrence of a death and its cause through the health system (essentially the public health system) before information from death certificates, OFS or OCP. This is particularly so for public sector, where information about the patient's follow-up is easier to obtain than in the private sector, with which communication is mainly based on mails and willingness of the practitioners.

Some patients leave the canton of Geneva after their diagnosis with cancer, but return and die in Geneva. These individuals are recorded as dead in the OFS database. However, since no additional information on their disease was collected in the Geneva area, they are considered lost to follow up at the point of their departure by the Geneva Cancer Registry.

Statistical methods

We first examined the agreement between the official underlying cause of death and the reviewed underlying cause of death. We then evaluated the impact of such disagreement on the cause-specific survival estimates.

We used the Kappa statistic to compare concordance between the two cause-of-death variables for all patients who had died (N = 5,065). The Kappa statistic corrects for agreement expected by chance alone. Its values range from 0 to 1; 0 represents no agreement whereas 1 is perfect agreement. We stratified the analysis according to age at diagnosis, age at death, period of diagnosis, period of death, social class, stage, treatment received, sector of care and place of death. Age at diagnosis and age at death were coded into 5 categories (0–49, 50–59, 60–69, 70–79 and 80 and over), whilst four periods were used for the temporal analysis of diagnosis and death (1970–79, 1980–89, 1990–99, 2000–09). Social class was based on the patient's last job or, if missing, on the patient's partner's job. It was divided in four categories (high, medium, low and unknown) [12]. Stage followed the TNM classification [13] with 5 subgroups (stage I, stage II, stage III, stage IV, unknown). We distinguished 5 categories for the treatment each patient received: surgery only, surgery plus adjuvant therapy, hormonal treatment, others (including a mix of different palliative therapies), and an absence of treatment. Only treatments received during the first six months after diagnosis are recorded by the registry according to the IARC rules [14]. Sector of care was defined as private or public sector. We also defined 5 categories of place of death: public hospital, retirement home, private hospital, patient's home and unknown.

We used variance-weighted least-squares regression to evaluate trends in the Kappa values for sub-groups [15].

Table 1 Baseline characteristics of the cohort of female breast cancer patients diagnosed in Geneva between 1970 and 2009

	Overall		Deceased	
	N	%	N	%
Age at diagnosis (mean, SD)	61,5 (0,14)		66,8 (0,21)	
Age groups				
0-50	2'422	23.0	749	14.8
50-59	2'469	23.4	831	16.4
60-69	2'383	22.6	1'055	20.8
70-79	1'949	18.5	1'317	26.0
80+	1'311	12.5	1'113	22.0
Period of diagnosis				
1970-79	1'890	17.9	1'576	31.1
1980-89	2'192	20.8	1'552	30.6
1990-99	2'896	27.5	1'302	25.7
2000-09	3'556	33.8	635	12.5
Socioeconomic status				
High	1'620	15.4	597	11.8
Middle	5'092	48.3	2'171	42.9
Low	2'390	22.7	1'484	29.3
Unknown	1'432	13.6	813	16.1
Stage				
I	3'434	32.6	1'014	20.0
II	4'355	41.3	2'044	40.4
III	1'219	11.6	807	15.9
IV	585	5.6	490	9.7
Unknown	941	8.9	710	14.0
Treatment				
Surgery only	1'890	17.9	1'252	24.7
Surgery + adjuvant	7'340	69.7	2'693	53.2
No treatment	473	4.5	421	8.3
Hormones only	547	5.2	448	8.8
Others	284	2.7	251	5.0
Sector of care				
Private	5'122	48.6	2'028	40.0
Public	5'412	51.4	3'037	60.0
Total	10'534	100.0	5'065	100.0

We used logistic regression to evaluate the odds of disagreement between the official and revised cause of death, associated with each of the factors listed above.

We also examined the concordance between the official and the revised cause of death as a function of time since diagnosis: patients who died within five years after diagnosis, patients who died after 5 years but before 10 years of follow-up and patients who died after 10 but before 15 years of follow-up. Because of small numbers, patients dying more than 15 years after their diagnosis were not considered.

To estimate the impact of discordance upon cause-specific survival, we derived Kaplan Meier cause-specific survival curves for the whole cohort (N = 10,534) using both official and revised cause of death. In cause-specific survival analyses, patients are classified as presenting the event if they are recorded as dying from their cancer while those who die from other causes are censored at the date of their death. We performed subgroup survival analysis by age group, period of diagnosis, stage of the disease and treatment.

Results

The cohort consisted of 10,534 women (mean age 61.5 years) diagnosed between 1970 and 2009. Nearly half belonged to the middle social class groups (Table 1). About three quarters of the women were diagnosed at early stage of disease (stage I and II). Almost 90% underwent surgery, associated with adjuvant treatments such as radiotherapy (63%), hormones (44%) and chemotherapy (33%; data not shown).

Among the 5,065 women who have died, the official and the revised underlying cause of death were identical for 4,620 patients (91%) (Table 2). 254 cases (5%) were recorded as dying of breast cancer according to their death certificate but as dying from other causes in the revised data. Among these women, the cause of death was mostly recoded to heart diseases (48%) and other malignant tumours (20%). Conversely, 191 cases (3.8%) were recorded as dying from other causes according to their death certificate but as dying from breast cancer in the revised data. Among these women, the main causes of death reported on their original death certificates were other malignant tumours (40%) or an imprecise code (19%) (Table 2). The overall value of the kappa test was 0.82 (p-value < 0.001).

Unadjusted concordance varied greatly between subgroups (Table 3). The concordance was significantly lower with increasing age, from 0.87 for ages 0–49 to 0.74 for ages 80+ (p-value for trend test = 0.008). Similar age-related trends, though not significant, were found among the three subpopulations defined by time since diagnosis. These age-related patterns were much less marked for age at death. Concordance was comparable in all four periods of diagnosis although it tended to be lower in the earlier periods. Concordance was greater for early stage of disease (stage I and II) compared to advanced stage (III and IV), from 0.84 for stage I to 0.63 for stage IV (p-value for trend <0.001). However, the concordance between the two underlying causes of death for women with missing stage (about 14%) tended to be higher than those for stage IV (and stage III). If these records corresponded to advanced diseases, as it is often the case, this stage-related pattern could be greatly attenuated. This pattern was more marked for patients deceased within the first five years after diagnosis. A clear pattern was found according to the type of treatment with higher concordance for complete, with curative intent, treatment (0.83), intermediate concordance for palliative treatment (0.73) and lower concordance for non-treated patients (0.63). This pattern was mostly found among patients who died within five years since

Table 2 Cause of death among women diagnosed with breast cancer in Geneva between 1970 and 2009: effect of reclassification of the official underlying cause of death by the Geneva Cancer Registry

Cause of death			**5'065**		
Concordant			**4'620**		
Breast cancer			*2'508*		
Other cause			*2'112*		
Discordant			**445**		
Distribution of discordant cases					
Revised cause of death	Breast cancer as the official cause of death		**Official cause of death**	Breast cancer as the revised cause of death	
	N	%		N	%
Other tumour	50	19.7	Other tumour	77	40.3
Heart disease	121	47.6	Heart disease	35	18.3
Imprecise code	11	4.3	Imprecise code	37	19.4
Other	72	28.4	Other	42	22.0
	254	100.0		191	100.0

Table 3 Concordance by subgroups between the official underlying cause of death and the revised underlying cause of death for women diagnosed with breast cancer in Geneva between 1970 and 2009

	All data				Between 0 and 4 years of follow-up				Between 5 and 9 years of follow-up				Between 10 and 14 years of follow-up			
	N	%	Kappa	SD	N	%	Kappa	SD	N	%	Kappa	SD	N	%	Kappa	SD
Overall	5,065	100.0	0.82	0.01	2,497	100.0	0.76	0.02	1,275	100.0	0.86	0.03	626	100.0	0.83	0.04
Age at diagnosis																
0-49	749	14.8	0.87	0.04	317	12.7	0.76	0.06	210	16.5	0.91	0.07	94	15.0	0.92	0.10
50-59	831	16.4	0.87	0.03	380	15.2	0.79	0.05	210	16.5	0.81	0.07	87	13.9	0.90	0.11
60-69	1,055	20.8	0.82	0.03	442	17.7	0.75	0.05	239	18.7	0.87	0.06	146	23.3	0.76	0.08
70-79	1,317	26.0	0.81	0.03	600	24.0	0.73	0.04	339	26.6	0.82	0.05	237	37.9	0.79	0.06
80+	1,113	22.0	0.74	0.03	758	30.1	0.71	0.04	277	21.7	0.79	0.06	62	9.9	0.63	0.12
Trend test				*p = 0.008*				*p = 0.394*				*p = 0.222*				*p = 0.105*
Age at death																
0-49	353	7.0	0.80	0.05	248	9.9	0.76	0.06	94	7.4	0.92	0.10	8	1.3		N/A
50-59	593	11.7	0.86	0.04	350	14.0	0.80	0.05	160	12.6	0.88	0.08	64	10.2	0.96	0.12
60-69	813	16.1	0.79	0.04	427	17.1	0.73	0.05	222	17.4	0.83	0.07	87	13.9	0.86	0.11
70-79	1,105	21.8	0.81	0.03	580	23.2	0.73	0.04	270	21.2	0.85	0.06	130	20.8	0.86	0.09
80+	2,201	43.5	0.77	0.02	892	35.7	0.72	0.03	529	41.5	0.81	0.04	337	53.8	0.73	0.05
Trend test				*p = 0.339*				*p = 0.405*				*p = 0.256*				N/A
Period of diagnosis																
1970-79	1,576	31.1	0.80	0.03	695	27.8	0.72	0.04	354	27.8	0.81	0.05	198	31.6	0.75	0.07
1980-89	1,552	30.6	0.80	0.03	686	27.5	0.69	0.04	387	30.5	0.84	0.05	206	32.9	0.82	0.07
1990-99	1,302	25.7	0.86	0.03	654	26.2	0.81	0.04	366	28.7	0.88	0.05	217	34.7	0.91	0.07
2000-09	635	12.5	0.84	0.04	462	18.5	0.81	0.05	168	13.2	0.90	0.07	5	0.8		N/A
Trend test				*p = 0.143*				*p = 0.036*				*p = 0.229*				N/A
Social Class																
High	597	11.8	0.81	0.04	281	11.3	0.76	0.06	144	11.3	0.86	0.08	77	12.3	0.77	0.11
Medium	2,171	42.9	0.85	0.02	1,034	41.4	0.78	0.03	553	43.3	0.89	0.04	290	46.3	0.86	0.06
Low	1,484	29.3	0.80	0.03	735	29.4	0.71	0.04	363	28.5	0.82	0.05	177	28.3	0.86	0.07
Unknown	813	16.1	0.79	0.04	447	17.9	0.76	0.05	215	16.9	0.85	0.07	82	13.1	0.73	0.11
Trend test[¥]				*p = 0.569*				*p = 0.297*				*p = 0.492*				*p = 0.556*
Stage																
Stage I	1,014	20.0	0.84	0.03	280	11.2	0.78	0.06	305	23.9	0.85	0.06	198	31.6	0.84	0.07
Stage II	2,044	40.4	0.83	0.02	901	36.1	0.77	0.03	564	44.2	0.87	0.04	286	46.7	0.84	0.06
Stage III	807	15.9	0.77	0.04	538	21.6	0.74	0.04	175	13.7	0.80	0.08	58	9.3	0.76	0.13
Stage IV	490	9.7	0.63	0.04	427	17.1	0.55	0.05	50	3.9	0.95	0.14	8	1.3	0.38	0.28
Unknown	710	14.0	0.79	0.04	351	14.1	0.70	0.05	181	14.2	0.82	0.07	76	12.1	0.84	0.11
Trend test[¥]				*p = 0.000*				*p = 0.000*				*p = 0.963*				*p = 0.272*
Treatment																
Surgery only	1,252	24.7	0.83	0.03	436	17.5	0.77	0.05	323	25.3	0.85	0.06	216	34.5	0.82	0.07
Surg + adj.	2,693	53.2	0.86	0.02	1,186	47.5	0.82	0.03	766	60.1	0.87	0.04	367	58.6	0.85	0.05
No treatment	421	8.3	0.63	0.05	317	12.7	0.61	0.06	69	5.4	0.79	0.12	23	3.7	0.47	0.21
Hormones	448	8.9	0.73	0.05	345	13.8	0.70	0.05	88	6.9	0.81	0.11	14	2.2	0.86	0.26
Others	251	5.0	0.65	0.06	213	8.5	0.60	0.07	29	2.3	0.87	0.18	6	1.0	0.57	0.37
Sector of care																
Private	2,028	40.0	0.86	0.02	870	34.8	0.81	0.03	550	43.1	0.86	0.04	286	45.7	0.87	0.06

Table 3 Concordance by subgroups between the official underlying cause of death and the revised underlying cause of death for women diagnosed with breast cancer in Geneva between 1970 and 2009 *(Continued)*

Public	3,037	60.0	0.80	0.02	1,627	65.2	0.73	0.02	725	56.9	0.86	0.04	340	54.3	0.79	0.05
Period of death																
1970-79	630	12.4	0.70	0.04	548	21.9	0.71	0.04	82	6.4	0.67	0.11	-	-		
1980-89	1,196	23.6	0.74	0.03	658	26.4	0.68	0.04	362	28.4	0.84	0.05	149	23.8		N/A
1990-99	1,483	29.3	0.83	0.03	694	27.8	0.78	0.04	377	29.6	0.87	0.05	211	33.7		
2000-09	1,756	34.7	0.88	0.02	597	23.9	0.84	0.05	454	35.6	0.88	0.05	266	42.5		
Trend test				*p = 0.000*				*p = 0.007*				*p = 0.140*				N/A
Place of death																
Public hospital	2,845	56.2	0.76	0.02	1,573	63.0	0.69	0.03	677	53.1	0.82	0.04	315	50.3	0.78	0.06
Retirement home	1,291	25.5	0.83	0.03	570	22.8	0.77	0.04	328	25.7	0.88	0.06	172	27.5	0.79	0.08
Private hospital	150	3.0	0.85	0.08	55	2.2	0.74	0.13	47	3.7	0.77	0.14	24	3.8	1.00	0.20
Home	374	7.4	0.91	0.05	143	5.7	0.94	0.08	113	8.9	0.87	0.09	54	8.3	0.90	0.14
Others	263	5.2	0.96	0.06	122	4.9	0.96	0.09	70	5.5	0.95	0.12	38	6.1	0.92	0.16
Missing	142	2.8	0.92	0.08	34	1.4	0.94	0.17	40	3.1	0.83	0.16	25	4.0	1.00	0.20

[y]Trend test performed without the missing data.

diagnosis. We found no association between social class and concordance, but a higher concordance for patients who were monitored (0.86) or who have died (0.85) in the private sector than for those in the public sector (0.80 and 0.76, respectively).

Unadjusted odds ratios of disagreement between the official and the revised underlying causes of death are presented in Table 4 for the overall cohort and for the three subcohorts defined by length of follow-up. The odds of disagreement increased significantly with age at diagnosis (as continuous variable) for all the patients (OR 1.03, 95% CI [1.02; 1.03]) and for the three subcohorts. We observed the same trend when using age at death as continuous variable (OR 1.02, 95% CI [1.01; 1.02] for all patients). Period of diagnosis was not significantly associated with disagreement but we did observe a significant decreasing trend for period of death as a continuous variable for all patients and the subcohorts (OR: 0.97, 95% CI: [0.96-0.98] for all patients). We did not find a significant trend for stage of the disease when considering all patients or the subcohorts defined by follow-up. Patients treated palliatively had significantly higher odds of disagreement (OR: 2.50, 95% CI [1.82; 3.43] for non-treated, 1.76 95% CI [1.26; 2.46] for patients treated with hormones and 1.34, 95% CI [0.86; 2.1] for other palliative treatment). The same trend was observed for the three subcohorts although not statistically significant. Patients treated in the public sector also had a higher risk of disagreement (OR: 1.47, 95% CI [1.20; 1.81] for all patients) as well as those who died in a public hospital. We did not observe differences by social class. We were unable to perform a logistic regression for the subcohort defined by a follow-up time between 10 and 15 years because of the small number of observations (<10) for several variables.

Figure 1 presents the breast cause-specific survival curves up to 20 years since diagnosis using the two different cause-of-death variables, for all breast cancer patients regardless their final vital status. The survival curves matched almost perfectly, with a difference in 20-year survival lower than 1%. The estimation of proportion of patients alive after twenty years of follow-up when using the official cause of death was 60.51%, 95% CI [59.11; 61.89] and 61.26, 95% CI [59.85; 62.64] when using the revised cause of death.

We compared cause-specific survival curves estimated with the revised and official underlying cause of death for selected subgroups (Figure 2). We estimated and presented results only if 10 women were remaining in the exposed group and/or the difference between the two curves was larger than 1%. Among patients aged 70–79 the survival at 20-year was 53.9% (95% CI [50.0; 57.6]) when using the revised cause of death and 51.2% (95% CI [47.3; 54.9]) with the official cause of death. The 20-year survival was greater when using the revised cause of death among the two first period of diagnosis. 1.7% and 1.6% difference for 1970–79 and 1980–89 respectively. We also observed a difference for patients treated with surgery. The 20-year survival was 65.2%, 95% CI [62.4; 68.0], based on the revised cause of death and 63.0%, 95% CI [60.1; 65.8] when using only death certificates.

A difference was already present at 10 years for several subgroups. Among patients with no treatment, the estimation was larger for reviewed cause of death with 3.5% difference. Among patients with hormonal therapy, the survival was 4.3% higher, 37.0%, 95% CI [29.5; 44.6] for

Table 4 Univariable logistic regression describing the disagreement by subgroups between the official underlying cause of death and the revised underlying cause of death for women diagnosed with breast cancer in Geneva between 1970 and 2009

	All data				Between 0 and 4 years of follow-up				Between 5 and 9 years of follow-up				Between 10 and 14 years of follow-up			
	N	%	OR	95% CI	N	%	OR	95% CI	N	%	OR	95% CI	N	%	OR	95% CI
Age (continuous)	5,065	100.0	1.03	[1.02;1.03]	2,497	100.0	1.02	[1.02;1.03]	1,275	100.0	1.02	[1.01;1.04]	626	100.0	1.03	[1.00;1.06]
Age at diagnosis																
60-69	1,055	20.8	1		442	17.7	1		239	18.7	1		146	23.3	1	
0-49	749	14.8	0.54	[0.37;0.81]	317	12.7	0.63	[0.37;1.07]	210	16.5	0.44	[0.17;1.15]	94	15.0	0.25	[0.71;0.88]
50-59	831	16.4	0.64	[0.45;0.92]	380	15.2	0.62	[0.37;1.03]	210	16.5	1.07	[0.50;2.27]	87	13.9	0.37	[0.12;1.12]
70-79	1,317	26.0	1.10	[0.83;1.46]	600	24.0	1.26	[0.85;1.86]	339	26.6	1.40	[0.73;2.67]	237	37.9	0.66	[0.33;1.32]
80+	1,113	22.0	1.54	[1.17;2.04]	758	30.1	1.48	[1.02;2.14]	277	21.7	1.48	[0.76;2.88]	62	9.9	1.12	[0.46;2.76]
Years (continuous)			0.99	[0.98;1.00]			0.99	[0.98;1.00]			0.97	[0.95;1.00]			0.94	[0.91;0.97]
Calendar period																
70-79	1,576	31.1	1		695	27.8	1		354	27.8	1		198	31.6		
80-89	1,552	30.6	1.00	[0.79;1.27]	686	27.5	1.31	[0.95;1.80]	387	30.5	0.84	[0.50;1.43]	206	32.9		N/A
90-99	1,302	25.7	0.69	[0.53;0.91]	654	26.2	0.80	[0.56;1.14]	366	28.7	0.63	[0.36;1.13]	217	34.7		
00-09	635	12.5	0.81	[0.58;1.13]	462	18.5	0.82	[0.56;1.22]	168	13.2	0.52	[0.23;1.16]	5	0.8		
Social Class																
High	597	11.8	1		281	11.3	1		144	11.3	1		77	12.3	1	
Medium	2,171	42.9	0.77	[0.56;1.06]	1,034	41.4	0.87	[0.56;1.33]	553	43.3	0.80	[0.38;1.66]	290	46.3	0.56	[0.24;1.28]
Low	1,484	29.3	1.04	[0.75;1.44]	735	29.4	1.24	[0.80;1.92]	363	28.5	1.30	[0.62;2.71]	177	28.3	0.55	[0.22;1.36]
Unknown	813	16.1	1.11	[0.78;1.59]	447	17.9	1.13	[0.70;1.81]	215	16.9	1.08	[0.47;2.45]	82	13.1	1.05	[0.40;2.74]
Stage																
Stage I	1,014	20.0	1		280	11.2	1		305	23.9	1		198	31.6	1	
Stage II	2,044	40.4	1.24	[0.93;1.66]	901	36.1	1.06	[0.68;1.64]	564	44.2	0.92	[0.53;1.61]	286	46.7	1.24	[0.61;2.52]
Stage III	807	15.9	1.48	[1.06;2.07]	538	21.6	1.03	[0.64;1.64]	175	13.7	1.36	[0.69;2.68]	58	9.3	1.95	[0.74;5.15]
Stage IV	490	9.7	1.25	[0.84;1.85]	427	17.1	0.87	[0.52;1.44]	50	3.9	0.28	[0.04;2.10]	8	1.3	4.74	[0.87;25.87]
Unknown	710	14.0	1.59	[1.13;2.23]	351	14.1	1.51	[0.93;2.44]	181	14.2	1.22	[0.61;2.44]	76	12.1	1.22	[0.45;3.33]
Treatment																
Surgery only	1,252	24.7	1		436	17.5	1		323	25.3	1		216	34.5	1	
Surg + adj.	2,693	53.2	0.78	[0.61;1.01]	1,186	47.5	0.58	[0.40;0.84]	766	60.1	0.85	[0.51;1.41]	367	58.6	0.99	[0.52;1.89]
No treatment	421	8.3	2.50	[1.82;3.43]	317	12.7	1.84	[1.23;2.75]	69	5.4	1.41	[0.58;3.41]	23	3.7	4.41	[1.53;12.75]
Hormones	448	8.9	1.76	[1.26;2.46]	345	13.8	1.34	[0.88;2.03]	88	6.9	1.25	[0.54;2.88]	14	2.2	0.96	[0.12;7.83]
Others	251	5.0	1.34	[0.86;2.10]	213	8.5	1.00	[0.60;1.67]	29	2.3	0.44	[0.06;3.41]	6	1.0	2.50	[0.28;22.71]

Table 4 Univariable logistic regression describing the disagreement by subgroups between the official underlying cause of death and the revised underlying cause of death for women diagnosed with breast cancer in Geneva between 1970 and 2009 *(Continued)*

Period of death																
continuous			0.97	[0.96;0.98]			0.98	[0.97;1.00]			0.98	[0.96;1.00]			0.94	[0.90;0.97]
70-79	630	12.4	1		548	21.9	1		82	6.4	1					
80-89	1,196	23.6	1.01	[0.75;1.36]	658	26.4	1.18	[0.84;1.66]	362	28.4	0.60	[0.28;1.30]	149	23.8		N/A
90-99	1,483	29.3	0.69	[0.51;0.93]	694	27.8	0.86	[0.6;1.23]	377	29.6	0.51	[0.24;1.11]	211	33.7		
00-10	1,756	34.7	0.45	[0.33;0.61]	597	23.9	0.65	[0.44;0.96]	454	35.6	0.44	[0.20;0.95]	266	42.5		
Age at death																
continuous			1.02	[1.01;1.02]			1.02	[1.01;1.03]			1.02	[1.01;1.04]			1.03	[1.00;1.05]
60-69	813	16.1	1		427	17.1	1		222	17.4	1		87	13.9	1	
0-49	353	7.0	0.65	[0.39;1.08]	248	9.9	0.66	[0.37;1.18]	94	7.4	0.30	[0.67;1.34]	8	1.3	2.34	[0.24;22.94]
50-59	593	11.7	0.49	[0.31;0.79]	350	14.0	0.51	[0.29;0.90]	160	12.6	0.54	[0.20;1.42]	64	10.2	0.26	[0.30;2.28]
70-79	1,105	21.8	1.13	[0.82;1.56]	580	23.2	1.25	[0.83;1.86]	270	21.2	1.10	[0.55;2.21]	130	20.8	1.22	[0.39;3.77]
80+	2,201	43.5	1.23	[0.93;1.63]	892	35.7	1.46	[1.01;2.10]	529	41.5	1.31	[0.72;2.41]	337	53.8	1.90	[0.72;5.01]
Place of death																
Public hospital	2,845	56.2	1		1,573	63.0	1		677	53.1	1		315	50.3		
Retirement home	1,291	25.5	0.74	[0.58;0.93]	570	22.8	0.87	[0.64;1.17]	328	25.7	0.71	[0.41;1.22]	172	27.5		
Private hospital	150	3.0	0.59	[0.31;1.14]	55	2.2	0.54	[0.19;1.51]	47	3.7	1.37	[0.52;3.62]	24	3.8		N/A
Home	374	7.4	0.37	[0.22;0.62]	143	5.7	0.20	[0.07;0.54]	113	8.9	0.76	[0.34;1.72]	54	8.3		
Others	263	5.2	0.13	[0.05;0.35]	122	4.9	0.11	[0.03;0.47]	70	5.5	0.17	[0.02;1.23]	38	6.1		
Missing	142	2.8	0.30	[0.12;0.75]	34	1.4	0.21	[0.03;1.53]	40	3.1	0.94	[0.28;3.13]	25	4.0		
Sector of care																
Private	2,028	40.0	1		870	34.8	1		550	43.1	1		286	45.7	1	
Public	3,037	60.0	1.47	[1.20;1.81]	1,627	65.2	1.61	[1.21;2.14]	725	56.9	1.02	[0.66;1.58]	340	54.3	1.60	[0.88;2.91]

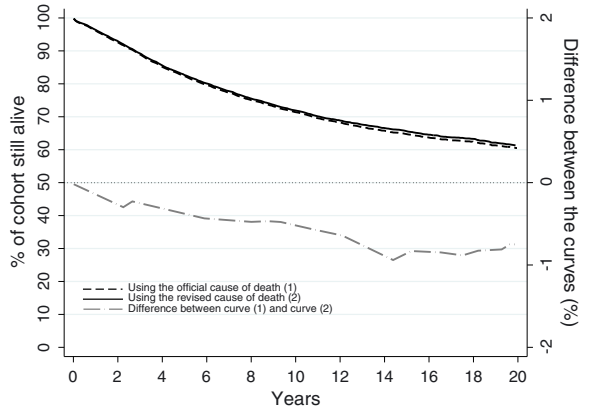


Figure 1 Up-to-20-year cancer-specific survival using 1) the cause of death based on death certificate only and 2) cause of death reviewed by registrars and the absolute difference between them: female breast cancer patients diagnosed between 1970 and 2009.

the reviewed cause of death vs. 32.7%, 95% CI [25.9; 39.6] when using the variable based only on death certificates. In the same way, the survival at 10-year was 5% higher for 80+ when using the reviewed cause of death (47.7%, 95% CI [43.1; 52.1] vs. 42.7%, 95% CI [38.4; 47.0]). Among patients with metastatic tumours, the difference was in the opposite direction: the estimation of 10-year survival was 1.5% higher when using the cause of death based on death certificates only, 14.7%, 95% CI [11.2; 18.7] vs.13.2%, 95% CI [10.0; 16.9] for the reviewed cause of death.

Discussion and conclusion

Survival statistics derived from routinely collected population-based cancer registry data are key means of reporting progress against cancer. In the Geneva Cancer Registry, in addition to the official underlying cause of death derived from the death certificate, registrars use all the available information in order to establish, where relevant, a revised underlying cause of death which allows evaluation of the accuracy of death certification.

This study describes both processes of recording the cause of death and shows their impact upon estimated survival rates from breast cancer.

The overall concordance between the official and the revised underlying cause of death was high. Differences were only present for 8.8% of the deceased patients representing 4.2% of the entire cohort. This is consistent with the study conducted by Goldoni et al. [11] in 2009 who reported 4.3% misclassification among their cohort. The official underlying cause of death was revised to breast cancer in 191 women (3.8% of those who have died) according to the cancer registry registrars; the underlying cause of death of these women had mainly been coded to other tumours. This could be explained by the presence of metastases that may have misled the certifying doctor about the location of the primary cancer and leads to differences in cause-specific survival estimation among metastatic patients (Figure 2).

On the other hand, most of the 254 women (5.0% of the patients who have died), coded as breast cancer deaths on the death certificates and considered as deaths from other causes from the registry, have been attributed to heart disease. Most of these women were elderly patients diagnosed during 1970–89. At that time the guidance for death certification among cancer registries was not to emphasize the cancer as a cause of death [16]. This might explain a tendency to recode the cause of death from cancer to heart diseases among elderly.

Our results based on Kappa statistic and on logistic regression showed that disagreement was greater among elderly women, patients with advanced disease and patients receiving palliative treatment. This suggests that less attention is given by doctors certifying death to the underlying cause of death for patients who are more likely to die. Concordance is also lower within the first five years after diagnosis, suggesting that more accurate information is available to the registrars assessing the true underlying cause of death during a shorter period of follow-up.

We also observed increasing concordance in successive calendar periods of death. Since this variable closely represents the year in which the review took place, several explanations may apply. First, the Geneva Cancer Registry may have less information in more recent times. This seems unlikely since more linkages have been set up over time with the health system in the canton, allowing a greater exchange of data. More likely, the accuracy of death certificates has improved over time which has led to more confidence in the official coding supplied on death certificates.

It is legitimate to ask why the reliability of cause of death reported on the death certificates may be questioned at all. It can be argued that the general practitioner responsible for the patient is the person most likely to be aware of the underlying cause of death insofar as they are aware of all the clinical information and also often know the patients personally. However, this advantage is not always capitalised on. Physicians are more likely to misclassify the cause of death than a trained registrar [4,10,17-19]. The general practitioner is not always concerned about the epidemiological information they are providing, and may not be aware of the international rules of WHO about the coding of the cause of death. Moreover, the general practitioner often receives the results of the autopsy after the death certificate has been issued and therefore does not take into account the report when certifying the death. The registrars of the

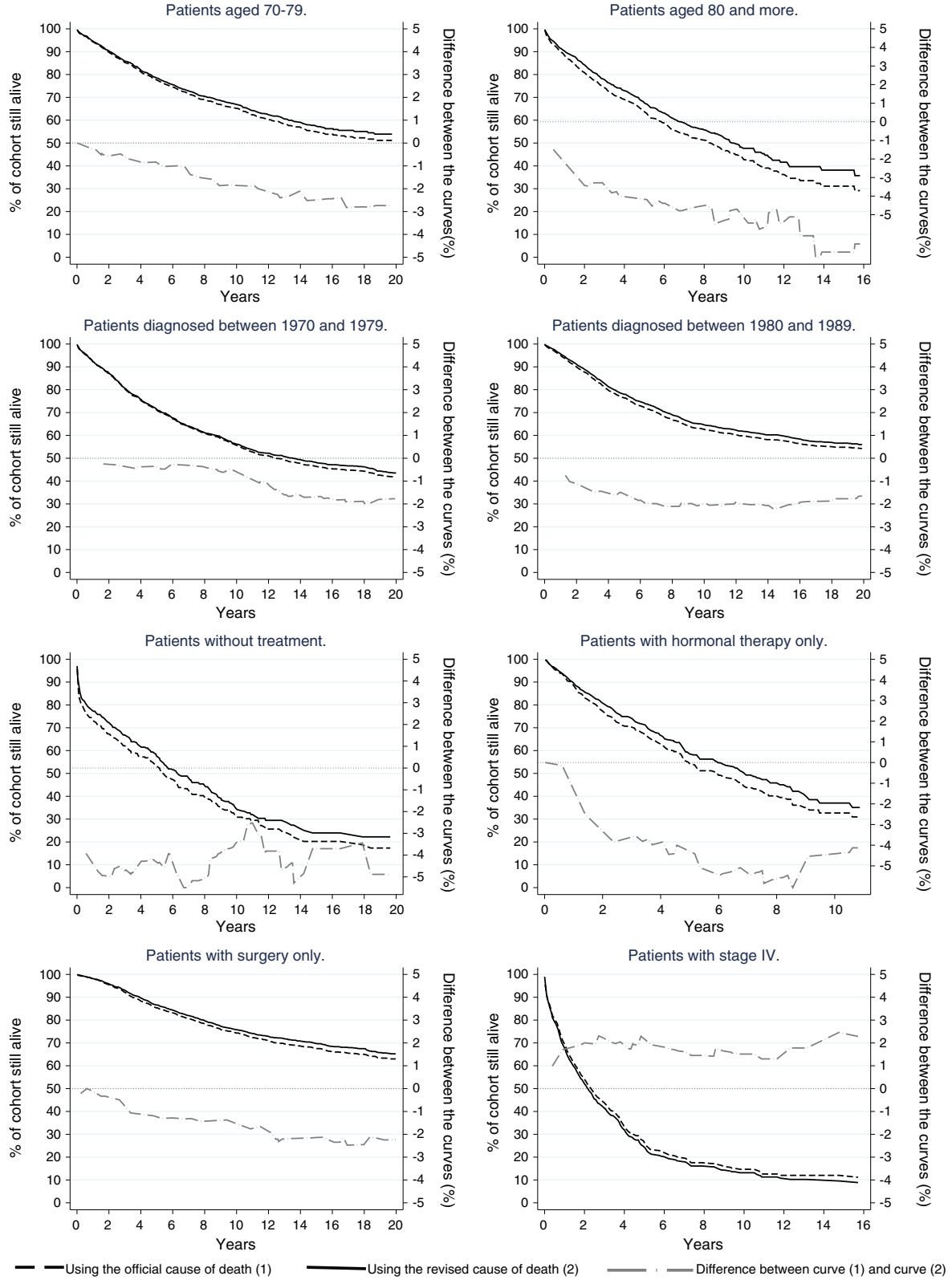


Figure 2 Up-to-20-year cancer-specific survival using 1) the cause of death based on death certificate only and 2) cause of death reviewed by registrars and the absolute difference between them: female breast cancer patients diagnosed between 1970 and 2009. Selected results by co-variables.

Geneva Cancer Registry, on the other hand, are able to access all pathological and histological information and/or the clinical information for most cases and review the cause of death only in the light of the autopsy reports. In addition, the registrars are more experienced with epidemiological data and its coding.

James [20] showed that coding the cause of death using death certificates only, in isolation from all other available information, led to biased interpretations of the cause of death. Our study tends to legitimate the process of verification that is performed in the Geneva Cancer Registry and induces that the resulting estimation of survival is more accurate.

Both methods aim to assign as best as possible the cause of death and none of them can be considered as the gold standard. We nevertheless consider the revised cause of death as more accurate insofar as additional information is available to experienced registrars but not necessarily to the practitioner completing the death certificate.

Overall, the revised underlying cause of death did not have a major impact on the cause-specific survival up to 20 years. However, important differences appeared in several subgroups suggesting that using the official underlying cause of death could lead to biased estimation of cause-specific survival in some populations.

The main limitation of this study relates to the proportion of women who have died for whom information other than the death certificate was available. The more information available, the more likely it is that we will be able to find discordance. High concordance reflects either lack of additional information available to correct the official cause of death, or that the death certificates define the cause of death fairly well. However, among our cohort of 5,062 deceased patients, a high percentage was monitored and/or passed away in the public sector of care, where access to information about cause of death is more readily available. We therefore assume that information enabling review of the underlying cause of death was available for the great majority of women who had died and that the overall high concordance between official and revised underlying cause of death is real.

Moreover, the number of deaths in the cohort influences the discordance. The more deaths, the more likely it is to find differences between the two causes-of-death and then the concordance. This state is confirmed in our study with a higher discordance among elderly. Breast cancer is not the more lethal cancer and our results can certainly not be generalised to other tumour localisations.

The Geneva Cancer registry data represent a unique opportunity to review the accuracy of the cause of death recorded on a death certificate by comparing it to all the available information in the health system. We observed that the overall concordance with the cause of death found on the death certificates is fairly high. More particularly, the impact on estimates of cause-specific survival is very small overall, although analyses in subgroups show larger differences, suggesting that misclassification of the underlying cause of death could lead to biased estimation of differences or trends in cause-specific survival.

Competing interest

There are no conflicts of interest to declare.

Authors' contributions

All authors contributed to the manuscript. RS conducted the analysis and the writing under the supervision of LW and BR. BR, LW and ER all reviewed the paper and made final corrections. All authors read and approved the final version of the manuscript.

Acknowledgements

This collaborative work was supported by the Geneva Cancer Registry and the Cancer Survival Group at London School of Hygiene and Tropical Medicine. We should like to thank Massimo Usel, Gerald Fioretta, Isabelle Neyroud and all collaborators at the Geneva Cancer Registry as well as Michel Coleman and all collaborators at the London School of Hygiene and Tropical Medicine for their support and encouragements.

Author details

[1]Geneva Cancer Registry, Institute for Social and Preventive Medicine, University of Geneva, 55 Boulevard de la Cluse, Geneva, 1205, Switzerland. [2]Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [3]Centre for Cancer Control and Statistics, Osaka Medical Centre for Cancer and Cardiovascular diseases, Osaka, Japan.

Received: 3 June 2013 Accepted: 21 November 2013
Published: 27 December 2013

References

1. Perme MP, Stare J, Estève J: **On estimation in relative survival.** *Biometrics* 2012, **68**(1):113–120.
2. Percy C, Stanek E, Gloeckler L: **Accuracy of cancer death certificates and its effect on cancer mortality statistics.** *Am J Publ Health* 1981, **71**(3):242–250.
3. Hoel DG, Ron E, Carter R: **Influence of death certificate errors on cancer mortality trends.** *J Nat Cancer Inst* 1993, **85**(13):1063–1068.
4. Messite J, Stellman SD: **Accuracy of death certificate completion: the need for formalized physician training.** *JAMA* 1996, **275**(10):794–796.
5. Maudsley G, Williams EM: **"Inaccuracy' in death certification–where are we now?** *J Publ Health Med* 1996, **18**(1):59–66.
6. Lee PN: **Comparison of autopsy, clinical and death certificate diagnosis with particular reference to lung cancer. A review of the published data.** *APMIS Suppl* 1994, **45**:1–42.
7. Lloyd-Jones DM, Martin DO, Larson MG, Levy D: **Accuracy of death certificates for coding coronary heart disease as the cause of death.** *Ann Intern Med* 1998, **129**(12):1020–1026.
8. Newschaffer CJ, Otani K, Mcdonald MK, Penberthy LT: **Causes of death in elderly prostate cancer patients andin a comparison nonprostate cancer cohort.** *J Natl Cancer Inst*, **92**(8):613–621.
9. Albertsen P: **When is a death from prostate cancer Not a death.** *J Natl Cancer Inst* 2000, **92**(8):590–591.
10. Cambridge B, Cina SJ: **The accuracy of death certificate completion in a suburban community.** *Am J Forensic Med Pathol* 2010, **31**(3):232–235.
11. Goldoni CA, Bonora K, Ciatto S, Giovannetti L, Patriarca S, Sapino A, Sarti S, Puliti D, Paci E: **Misclassification of breast cancer as cause of death in a service screening area.** *Cancer Causes Control* 2009, **20**(5):533–538.
12. Rapiti E, Fioretta G, Schaffar R, Neyroud-Caspar I, Verkooijen HM, Schmidlin F, Miralbell R, Zanetti R, Bouchardy C: **Impact of socioeconomic status on prostate cancer diagnosis, treatment, and prognosis.** *Cancer* 2009, **115**(23):5556–5565.
13. Sellers AH: **"The clinical classification of malignant tumours: the TNM system".** *Can Med Assoc J* 1971, **105**(8):836.

14. Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG: *Cancer registration: principles and methods*. 95th edition. IARC scientific publications; 1991:1–288.
15. Grizzle JE, Starmer CF, Koch GG: **Analysis of categorical data by linear models.** *Biometrics* 1969, **25**(3):489–504.
16. Lutz J, Pury P, Fioretta G, Raymond L: **The impact of coding process on observed cancer mortality trends in Switzerland.** *Eur J Cancer Prev* 2003, **8**:77–81.
17. Smith Sehdev AE, Hutchins GM: **Problems with proper completion and accuracy of the cause-of-death statement.** *Arch Intern Med* 2001, **161**(2):277–284.
18. Lakkireddy DR, Gowda MS, Murray CW, Basarakodu KR, Vacek JL: **Death certificate completion: how well are physicians trained and are cardiovascular causes overstated?** *Am J Med* 2004, **117**(7):492–498.
19. Pritt BS, Hardin NJ, Richmond JA, Shapiro SL: **Death certification errors at an academic institution.** *Arch Pathol Lab Med* 2005, **129**(11):1476–1479.
20. James DS, Bull AD: **Information on death certificates: cause for concern?** *J Clin Pathol* 1996, **49**(3):213–216.

doi:10.1186/1471-2407-13-609
Cite this article as: Schaffar *et al.*: Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva cancer registry. *BMC Cancer* 2013 **13**:609.

[Chapter Two]

Comparison of the two data settings

I have shown that routinely recorded cause of death can be inaccurate and a source of bias in cancer survival estimates for some sub-groups in the cause-specific setting, and particularly for long-term survival. I therefore conclude the reviewed cause of death as the more accurate variable for estimating cancer survival in the cause-specific setting (Chapter 1).

However, not all cancer registries have access to accurate information about the underlying cause of death, making it difficult to estimate cancer survival using this approach. The alternative is the use of the relative-survival setting where this information is not required.

The purpose of this chapter is to complete the second aim of the thesis, which is to evaluate the less biased data setting for the estimation of long-term net survival when reviewed cause of death is known.

In order to achieve this aim, I define the following objectives:

- to derive up-to-date life tables for use in the estimation of net survival within the relative survival setting,
- to apply an inverse probability weighting (IPW) method to estimate net survival in the cause-specific setting
- to compare and contrast estimates of long-term net survival from breast cancer in each data setting using both routinely collected and validated data on cause of death and
- to assess whether these same findings apply to other anatomic localisations.

This chapter comprises copies of two published research articles, a full paper published in *Cancer Epidemiology* and a Short Communication published in

the *European Journal of Cancer,* alongside text, which describes the background to the objectives, summarises the approach and findings, and which specifies how these publications fulfil the aim and objectives.

Background

We distinguish two sources of bias when estimating net survival; the first one is related to the theoretical definition of the estimator itself and the second depends on the data setting used for its estimation.

Net survival and informative censoring

Net survival is the survival that would be observed in the hypothetical situation where the disease of interest, breast cancer in our case, is the only cause of death.

The main underlying assumption of net survival is that death from cancer is independent of background population mortality (the expected rate of death in the absence of breast cancer). In practice, however, this is very rarely the case because causes of death other than breast cancer can interfer with the cancer death itself [73–75]. This is described as a scenario of "competing risks" where the probability of a patient being censored and leaving the risk set is associated with the occurrence of the event of interest. In other words, patient characteristics tend to be associated with both deaths from the disease of interest (breast cancer) as well as deaths from other causes.

The association between covariables and withdrawal breaks the assumption of independence between the censoring process and the occurrence of the event. For instance, we know that old patients are more likely to die from other causes of death, due to co-morbidities, and are therefore more likely to be censored. This phenomenon is known as informative censoring. Without due consideration of informative censoring, estimates of net survival are biased.

Adjustment for informative censoring

Application in the relative survival setting

Historically, approaches defined by Ederer in 1961 [77], and in 1959 [78] (known as Ederer I and II respectively) as well as by Hakulinen in 1982 [79] were used to estimate survival from cancer in the relative survival setting. All three of these methods were considered, wrongly, to be estimating net survival. Actually, they estimated net survival only when the censoring process was entirely non-informative. This was demonstrated by Pohar-Perme *et al.* in 2011 [80] who showed that in practice an absence of informative censoring is almost never the case and that all of these three commonly applied methods, as well as survival estimates derived in the cause-specific setting, were biased in the presence of informative censoring.

To address the problem of informative censoring within the relative survival setting, Pohar-Perme [80] proposed a correction based on inverse probability weighting. The weight *w* is put on both cumulative overall mortality rate $\widehat{\Lambda}_o$ and cumulative expected mortality rate $\widehat{\Lambda}_e$ in order to derive the cumulative excess mortality rate $\widehat{\Lambda}_{excess}$. We have,

$$\widehat{\Lambda}_{excess}(t) = \widehat{\Lambda}_o(t) - \widehat{\Lambda}_e(t)$$

$$\widehat{\Lambda}_{e\ xcess}(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i{}^w(u) d\Lambda_{e_i}(u)}{Y^w(u)},$$

where $N^w(t) = \sum_{i=1}^n N_i^w(t) = \sum_{i=1}^n \frac{N_i(t)}{S_{e_i}(t)}$ and $Y^w(t) = \sum_{i=1}^n Y_i^w(t) = \sum_{i=1}^n \frac{Y_i(t)}{S_{e_i}(t)}$. $S_{e_i}(t)$ represents the expected survival for each individual i, derived from the general population life table. This approach therefore increases the weight of the

remaining risk set by dividing it by the expected probability of survival and thus enables the mimicking of the cohort that would have been observed without the withdrawals, which occur due to competing events (death from causes other than breast cancer). Consequently, the bias of informative censoring is taken into account.

Pohar-Perme is now considered the Gold Standard for the estimation of net survival within the relative survival setting and has been applied in Paper 2 as well as in the Short Communication.

Application in the cause-specific setting

Robins [81] and Satten [82] were the first to propose a solution to avoid the bias associated with informative censoring within the cause-specific setting. More precisely, Robins' idea was to "develop statistical methods that can be used to adjust for non-random non-compliance and dependent censoring in randomized control trial" [81]. Satten suggested the estimation of survival as a missing data issue as patients were censored and therefore likely not to present the event of interest. The solution was therefore the use of an "inverse-probability-of-censoring" weighting. For this purpose, they derived survival using the Nelson-Aalen estimator which considers two counting processes. The first, N, represents the number of events. The second, Y, corresponds to the number of individuals at risk. To tackle informative censoring, both processes are weighted, for each individual i, by the inverse of their probability of being censored $S_{ci}(t)$. For an individual i ($i = 1, \dots, n$), the cumulative net survival $\widehat{\Lambda}_{cs}^{w}(t)$ weighted by w, is:

$$\widehat{\Lambda}_{cs}^{w}(t) = \int_0^t \frac{dN_{cs}^{w}(u)}{Y^{w}(u)},$$

with $N_{cs}^{w}(t) = \sum_{i=1}^{n} N_{csi}^{w}(t) = \sum_{i=1}^{n} \frac{N_{csi}(t)}{S_{ci}(t)}$ and $Y^{w}(t) = \sum_{i=1}^{n} Y_{i}^{w}(t) = \sum_{i=1}^{n} \frac{Y_{i}(t)}{S_{ci}(t)}$. $N_{cs}^{w}(t)$ represents the number of deaths at time t, weighted by w. $Y^{w}(t)$ represents the number of individual at risk at time t, weighted by w. This enables survival to be estimated for the hypothetical cohort, which would have been observed without the withdrawals and therefore eliminates the bias due to informative censoring.

I used a similar strategy to create my own weights for the estimation of net survival in the cause-specific setting. I derived the weights using the cancer patient data and validated cause of death. I considered that the expected mortality of the cancer patients would be the same as the mortality rate from other causes of death than breast cancer amongst the cancer patients. I fitted a Poisson regression model to the cancer patient data where I considered death from a cause other than breast cancer as the event of interest. I adjusted on age at death and year of death. I used the model to derive expected mortality by age and year. I then used this set of rates to weight the breast-specific mortality hazard, in order to derive net survival estimates. I make use of this approach to derive my own weights when estimating net survival within the cause-specific setting in Paper 2 and in the Short Communication.

Two data settings, two data biases

Both cause-specific and relative-survival settings can be used to estimate net survival but, as well as being subject to bias as a result of informative censoring, survival estimates derived in both settings are also prone to important biases related to the data itself.

Data bias within the cause-specific setting

The main bias for the cause-specific setting is misclassification of cause of death. This issue is related to the accuracy of cause of death classification, and has been covered in detail in Chapter 1.

Data bias within the relative survival setting

Estimation of net survival in the relative survival setting is achieved by taking the ratio between the observed survival rate in a specified group of patients during a specified period of time, to the expected survival rate, which is the mortality rate observed in the population from which the patients are drawn. The major advantage of the relative survival setting in comparison to the cause-specific setting is that information about cause of death is not needed.

In the relative survival setting, the most important potential bias is the non-comparability between the cohort of cancer patients and the general population used for the estimation of the expected mortality rate. This is particularly important when a factor having an impact on mortality from other causes is distributed differently between the cancer group and the general population, that is, the population used for comparison does not provide an accurate estimation of the expected mortality of the patient group. Tobacco and lung cancer is an interesting illustration: patients with lung cancer are known to have much greater exposure to tobacco compared to the general population. Thus, their risk of dying from another smoking-related disease is considerably larger [83,84]. As a result, the expected mortality derived from the general population will be an under-estimate of their expected mortality and their net survival under-estimated. Ellis *et al.* attempted to adjust for this issue

using smoking-adjusted life tables [84]. Several other factors may act in this manner, including physical activity, obesity or diet [86–90].

In this respect, it is essential to derive the most accurate general population life table as possible.

Derivation of life tables

Life tables are demographic tools, which detail the probabilities of death amongst a given population by a set of covariates, usually age and sex. For the Geneva canton, mortality and population data are available at a very detailed level. The numbers of deaths and the population size (death and population counts) by single year of age and sex are accessible annually from 1970 to 2013. Such 'complete' (single year of age) life table data are attractive because of the level of detail they provide. However, the observed mortality rates tend to be quite variable especially in the context of a small population like Geneva. Consequently, it is unadvisable to use them directly in relative survival analyses. On the other hand, abridged life tables (which present probabilities for age groups) are much less prone to random variability but are not detailed enough since expected mortality rates only available by age group.

I followed the recommendation provided by Rachet *et al.* [91] that the most appropriate life table for use in the relative survival setting is both complete and smooth. They have demonstrated the use of a flexible Poisson regression model including splines to derive such life tables from routinely recorded death and population data. This model performs better than previously applied methods developed by Ewbank *et al.* [92] and Elandt-Johnson *et al.* [93] which are based on strong underlying assumptions about the distribution of deaths by age.

I modelled counts of death within the generalized linear model framework, using a Poisson error and log-link. Person-years at risk were used as the offset. I first considered the effect of age. More specifically, the flexible model defines:

$$\log(d_x) = \beta_0 + f(x) + \log(pyrs_x)$$

where x represents age in years, d_x denotes the age-specific death count, β_0 is the coefficient at baseline (i.e. the log of the mortality rate at the reference age), $f(x)$ denotes a restricted cubic spline function on age, and $pyrs_x$ denotes the age-specific person-years at risk. The model was implemented using the Stata command *mvrs* (multivariable regression splines) [94] in STATA. Splines are made up of piecewise polynomial functions joined at locations called knots.

As recommended by Rachet *et al.* [91], I tested six slightly different scenarios for the location of knots for the main effect of age upon the mortality rate. I then compared each of these scenarios and selected the best model according to AIC and BIC criteria. Then, I included the effect of calendar year upon the risk of death. I considered three hypotheses regarding the effect of year on mortality. First, I assumed that the change in mortality was constant over time (linear effect of year). Second, I allowed the change in mortality to be non-linear by the inclusion of a cubic spline. Finally, I modelled an interaction term between age and the non-linear function of year. This allowed mortality to change in a non-linear way for each separate single year of age. I compared these three models using AIC and BIC criteria in order to select the most appropriate one. The model presenting the interaction between age and year demonstrated the best fit and was retained. I then used this model to derive expected mortality rates for the general population by sex and single year of age for the all the years from 1970 up to 2013.

Results for women are illustrated in Figure 9. The observed (dots) and fitted (lines) mortality rates are presented by age bands. I observed large variability in mortality rates when using observed data, especially for middle-aged (30-65 years old) and the oldest individuals (more than 95), whereas fitted rates provide a more stable function for mortality. These fitted mortality rates were used for the estimation of net survival within the relative survival setting.

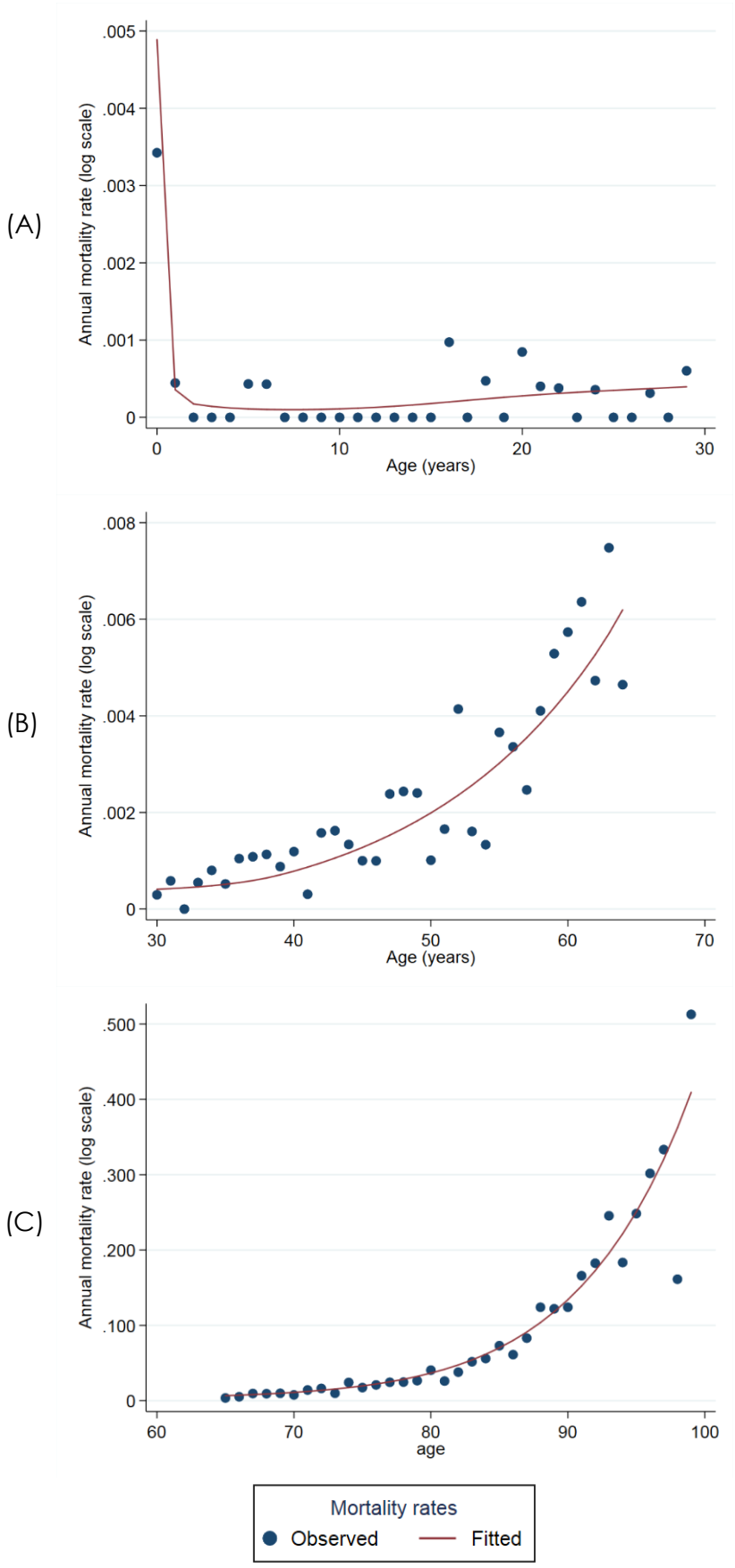


Figure 9: Observed and fitted age-specific mortality rates for the Geneva canton general population. Year 2000. (A) Females aged less than 35. (B) Females aged 35-64. (C) Females aged 65 and more.

Paper Two

Comparison of the two data settings for breast cancer.

Description

"**Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis**" was published in *Cancer Epidemiology* in 2015 [94].

In it, I compare the theoretically unbiased estimates of net survival available for each data setting. I implemented the weights inspired by Robins and Satten for the estimator derived (1) in the cause-specific setting using reviewed cause of death and (2) used the Pohar-Perme estimator for the relative survival setting. To evaluate the impact of the types of data biases between the settings I performed sensitivity analyses for (1) by randomly allocating a greater proportion of deaths to the cancer patients and for (2) by artificially increasing the expected mortality derived from the general population life table.

The first sensitivity analyses (1) focused on testing the extent of the possible data bias within the cause-specific setting. It is known that misclassification of the underlying cause of death is an issue in this context, in particular, deaths caused by the disease of interest but attributed to other diseases. I therefore artificially increased the number of deaths due to breast cancer, in order to mimic different levels of misclassification by randomly re-attributing the cause of death variable from non-breast cancer to breast cancer for 10, 15, 20 and 25 per cent of the deceased patients, and evaluated the impact of this upon net survival estimates. The estimation within the cause-specific setting would have been considered robust to this bias if different levels of misclassification led to

very similar estimations of net survival. Conversely, if a very small change in the proportion of deaths attributable to breast cancer induced a substantial change in the net survival estimates, this would constitute evidence that the cause-specific setting is sensitive to misclassification of the cause of death.

The second sensitivity analysis (2) addressed relative-survival setting. In this context, the most likely data bias is related to the non-comparability between the cancer patients and the population from which I extract the expected mortality rate. I therefore artificially modified the life table itself, increasing and stratifying the expected rate of death, in order to evaluate the impact of such a change upon the estimates of net survival. Had similar estimations of survival been obtained in this analysis I would have concluded that the relative survival setting is robust to non-comparability in the life tables. Conversely, the estimator would be found to be less reliable if a relatively small change in the expected rates of death led to a very different estimate of net survival.

Main results

Within this research, I show that using the cause-specific setting led to higher estimation of net survival compared to the relative survival setting. The absolute difference between the two estimators increased with time after diagnosis from 1% at one year to 10.8% at 20 years. It remained less than 3% during the first ten years of follow-up (2.4% at 10 years) and started to increase more dramatically from 13 years onwards. A possible explanation for this result is that relative-survival setting takes into account death indirectly due to breast cancer, whereas the cause-specific setting does not. Using the cause-specific setting means that an explicit decision has to be made regarding the allocation of the cause of death whether it was fully attributable to breast cancer or not.. For

some others, the cancer is a contributing factor and it is difficult to determine whether the death is entirely due to the disease of interest. The line between a cancer-specific death and a cancer-consequent death is usually a matter of judgement. In the relative survival setting, this issue does not apply insofar as the aim is to measure deaths that are in excess of what would be expected for the patients under study. Indirect deaths are therefore taken into account implicitly in the approach and could explain the differences in net survival estimates.

I also demonstrated that the relative survival setting is much more robust to data biases than the cause-specific setting when estimating long-term net survival for breast cancer patients. Indeed, a very small modification in the proportion of deaths due to breast cancer led to large variability in the estimation of net survival in the cause-specific setting. Conversely, estimation of net survival in the relative survival setting was much less sensitive to large changes in the expected mortality rates derived from the general population life table.

Conclusion

Our conclusion was that net survival derived using the cause-specific setting is very sensitive, and estimates are likely to be an over-estimation of the true net survival. On the other hand, the estimation of long-term net survival in the relative survival setting was more likely to be close to the true net survival estimation because it showed robustness to violations of the assumptions. These results were increasingly important as time since diagnosis increased. I therefore concluded the relative survival setting as the less biased method for estimating long-term survival from breast cancer.

Short communication

Extension to other localisations

Description

"**Estimation of net survival for cancer patients: Relative survival setting more robust to some assumption violations than cause-specific setting, a sensitivity analysis on empirical data.**" was published in the *European Journal of Cancer* in 2017 [95].

Paper 2 comprehensively compared the two data settings for breast cancer patients. Breast cancer, however, remains a particular cancer localisation. It is relatively unusual because survival is high but patients still exhibit excess mortality associated with their disease several decades after their diagnosis. This pattern of mortality related to the disease is rarely seen in cancers arising at other anatomic sites.

I was therefore interested in testing whether the results that were reported in Paper 2 were similar for cancers of different localisations, since my findings for Paper 2 have wide-ranging implications for the coding of cause of death and the estimation of net survival from cancer registry data. I tested three different cancers: lung cancer, a very aggressive disease with poor survival; colorectal cancer, less aggressive but affecting older people and presenting a moderate prognosis; and melanoma, a cancer arising more often among younger people and demonstrating favourable prognosis. With these three sites, I was thus able to evaluate these results for tumours with different lethality, different patterns of mortality and different age distributions.

Main results

I repeated the analyses conducted for Paper 2 to compare the results found for breast cancer using data on melanoma, lung and colorectal cancers. For colorectal cancers, 87% of the patients died, 511 of their cancer (59%). Among women diagnosed with lung cancer, 97% died. There were 483 deaths, 419 from lung cancer (87%). Among patients with melanoma, 46% died, 40 due to melanoma (29%). For breast cancer, 1700 patients died (68%) 844 from their cancer (50%).

We observed, consistent with our previous analyses, that net survival estimates using the cause-specific setting are higher than the estimates using the relative survival setting for every localisation. The absolute difference between the two estimators increased with time since diagnosis for all four cancers. For colorectal cancer, the difference widened from almost 2% at one year to over 7% at 20 years. For lung cancer, the difference increased sharply within the first two years after diagnosis (3% at two years) and moderately afterwards. There was no detectable difference during the first three years after diagnosis for melanoma, but it subsequently increased to more than 8% at 20 years after diagnosis.

Analyses in the cause-specific setting again highlighted the lack of robustness of the net survival estimators to re-allocation of the cause of death, irrespective of cancer site. The relative survival setting demonstrated much more stable net survival estimations for all cancer localisations when the expected rate of mortality was modified.

Conclusion

I concluded that the use of relative survival setting is recommended for this estimation of net survival, regardless of the cancer site. Even if not all localisations have been tested, I have demonstrated that results found for breast cancer were robust to change in lethality of the cancer and in the age distribution of the cancer patients.

Fulfilment of Aims and Objectives

In this Chapter, I have demonstrated that the relative survival setting is robust to large data change in the general population life table but that the estimation of net survival within the cause-specific setting is very sensitive to the allocation of the underlying cause of death. I have shown that this is true for breast cancer as well as for a range of other cancer sites.

True, underlying net survival is never perfectly estimated, even within the relative-survival setting. However, with informative censoring taken into account, my analyses have shown that the relative survival setting is less prone to bias and thus more likely to provide close estimates of the true net survival compared to the cause-specific setting, even in the presence of reviewed cause of death. Since, in chapter 1 I have shown that cause of death from standard certificates can be misclassified, it follows that the bias in cause-specific survival generated when using cause only from death certificates is likely to be greater than that observed in these analyses.

An important aspect of these results is that these data biases were especially important for long-term net survival, because of the increasing likelihood of misclassification of the underlying cause of death (as shown in Chapter 1) when fewer deaths occur. This is intuitive, since attributing cancer as the cause of death is more likely when death is temporally close, rather than a long time after diagnosis. The impact of these errors on survival estimates are thus exacerbated through time as they are effectively multiplied together.

It is important to note that both estimators were theoretically unbiased since weights were applied in both settings in order to take into account informative censoring. I used the weights proposed by Pohar-Perme in the relative-survival

setting and developed my own based on the work of Robins and Satten for the cause-specific setting. This represents the key strength of my research. I was therefore able to perform a valid comparison between the data settings, which has not been done previously.

Thus, I have so far shown that although there is evidence that validated death data affords an advantage in estimating survival for some sub-groups, particularly in the long-term, the cause-specific setting itself is more prone to bias in comparison to the relative survival setting. I conclude that because of the level and nature of this bias, the cause-specific setting should not be applied, and that the relative survival setting is the best way to measure long-term net survival, even when accurate cause of death information is available.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Robin Schaffar
Principal Supervisor	Dr Laura Woods
Thesis Title	Long-term net survival among women diagnosed with breast cancer: accuracy of its estimation and evaluation of its determinants

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	Cancer Epidemiology		
When was the work published?	2015		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. I planned and carried out the literature review and data analysis. I prepared all drafts of the paper. The co-authors provided input and feedback on the content data analysis and on the paper drafts prepared by me.

Student Signature: [redacted] **Date:** 09/03/2018

Supervisor Signature: [redacted] **Date:** 09/03/2018

Cancer Epidemiology 39 (2015) 465–472

Contents lists available at ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis

Robin Schaffar [a,b,*], Bernard Rachet [b], Aurélien Belot [b], Laura Woods [b]

[a] Geneva Cancer Registry, Global Health Institute, University of Geneva, Geneva, Switzerland
[b] Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

ARTICLE INFO

Article history:
Received 6 January 2015
Received in revised form 2 April 2015
Accepted 5 April 2015
Available online 20 April 2015

Keywords:
Net survival
Breast cancer
Cause-specific
Relative survival
Informative censoring

ABSTRACT

Background: Both cause-specific and relative survival settings can be used to estimate net survival, the survival that would be observed if the only possible underlying cause of death was the disease under study. Both resulting net survival estimators are biased by informative censoring and prone to biases related to the data settings within which each is derived. We took into account informative censoring to derive theoretically unbiased estimators and examine which of the two data settings was the most robust against incorrect assumptions in the data. *Patients and methods*: We identified 2489 women in the Geneva Cancer Registry, diagnosed with breast cancer between 1981 and 1991, and estimated net survival up to 20-years using both cause-specific and relative survival settings, by tackling the informative censoring with weights. To understand the possible origins of differences between the survival estimates, we performed sensitivity analyses within each setting. We evaluated the impact of misclassification of cause of death and of using inappropriate life tables on survival estimates. *Results*: Net survival was highest using the cause-specific setting, by 1% at one year and by up to around 11% twenty years after diagnosis. Differences between both sets of net survival estimates were eliminated after recoding between 15% and 20% of the non-specific deaths as breast cancer deaths. By contrast, a dramatic increase in the general population mortality rates was needed to see the survival estimates based on relative survival setting become closer to those derived from cause-specific setting. *Conclusion*: Net survival estimates derived using the cause-specific setting are very sensitive to misclassification of cause of death. Net survival estimates derived using the relative-survival setting were robust to large changes in expected mortality. The relative survival setting is recommended for estimation of long-term net survival among patients with breast cancer.

1. Introduction

Net survival is defined as the survival that would be observed if the only possible underlying cause of death was the disease under study [1]. This definition of survival probability is of particular interest since it is not influenced by changes in mortality from other causes and therefore allows accurate evaluation of survival from the disease, essential for cancer control.

Two main approaches have been developed to estimate net survival, each requiring different data settings and assumptions. First, the cause-specific approach, which requires a data setting with reliable individual information on the underlying cause of death. Thus, only deaths from the cancer under study are defined as events whilst others are censored. Second, the relative survival approach [2] compares the overall survival of a cohort of patients to that which they would have experienced if they had had the same mortality experience of the general population from which they were drawn. This approach requires a different data setting, where mortality data about the population from which the cancer

* Corresponding author at: Geneva Cancer Registry, Global Health Institute, University of Geneva, 55 Boulevard de la Cluse, 1205 Geneva, Switzerland. Tel.: +41 22 379 49 57.
E-mail address: robin.schaffar@unige.ch (R. Schaffar).

http://dx.doi.org/10.1016/j.canep.2015.04.001
1877-7821/

patients are drawn is available. Information about the cause of death is not required and we assume that the cancer-specific mortality included in the overall mortality is negligible compared to the overall mortality.

Both approaches are prone to a bias called *informative censoring* [3]. This is where the assumption of independence between the censoring process and the occurrence of the event (death) does not hold. For instance, an older patient is more likely to die from other causes than the disease under study than a younger patient. Thus, the older patients are more likely not to experience the death from cancer of interest simply because of their older age. The censoring process is therefore dependant on age and becomes informative. To take into account this bias, Robins [4] and Satten [5] proposed to weight the observed data by the inverse of the probability of not dropping out of the risk set, in order to find a cohort which would have been seen without the withdrawals. Pohar Perme [6] used this idea to propose an unbiased estimator of net survival within the relative survival setting.

As long as informative censoring is accounted for appropriately, both cause-specific and relative survival approaches derive theoretically unbiased estimators of net survival. However these estimators are prone to biases related to the data settings within which each is derived. These biases are independent of the method of estimation.

In the cause-specific data setting, what defines a cancer-related death versus a death from another cause is reliant upon the judgment of the person extracting the information and often prone to misclassification. Several studies have described this bias as being non-negligible [7–14]. For this reason the relative survival method has generally been preferred to estimate net survival with population-based data [15,16]. However, within the relative survival data setting non-comparability between the cohort of patients and the general population [17] life tables used can also lead to bias. If a factor is differently distributed between patients and the general population, the resulting expected mortality of the cohort will be incorrectly estimated [18]. For instance, patients with lung cancer are more often smokers compared with the general population. Their expected mortality is therefore under-estimated as they are more likely to die from other causes than the general population [19]. In the long term, this under-estimation may be balanced by the selection process over time of the more robust patients, who may die less than the general population [20]. This may impact net survival estimates [21]. Similarly, several factors can be associated with both cancer mortality and other diseases and lead to non-comparability between observed and expected mortality.

Our objective was to compare the two data settings, cause-specific and relative survival, when estimating long-term net survival. Both are subject to bias as described above; either misclassification of the cause of death or use of inappropriate life tables. We first derived theoretically unbiased estimators by using weights for both approaches, which took into account informative censoring. We then performed two sensitivity analyses in order to examine which of the two data settings was more robust against incorrect assumptions. We used each estimator as a reference for the other in order to evaluate the impact on the net survival estimates (Table 1).

We used data from the Geneva Cancer Registry which holds high quality data on cancer patients collected since 1970. This enabled us to evaluate the effect of these biases on long-term net survival. Furthermore, it afforded a privileged situation for estimating net survival within the cause-specific setting as information on cause of death had been independently verified.

2. Material and methods

2.1. Data

The data were provided by the Geneva Cancer Registry.

The Geneva Cancer Registry collects information on incident cancer cases from various sources, including hospitals, laboratories and private clinics, all requested to report new cancer cases. Trained registrars systematically extract information from the medical records and conduct further investigation in the case of missing key data. The registry regularly assesses survival, taking as the reference date the date of confirmation of diagnosis or the date of hospitalization (if it preceded the diagnosis and was related to the disease). In addition to passive follow-up (standard examination of death certificates and hospital records), active follow-up is performed yearly using the files of the Cantonal Population Office who maintain a register of the resident population. The cause of death is validated or revised from death certificates by registrars using all available clinical information. Autopsy reports, letter at death written by general practitioners and all patients' medical notes are used for the assessment of the revised cause of death. The treatment can therefore be considered as breast cancer death when information is found about it being part of the morbid events leading directly to death [22]. We included all women diagnosed with an invasive primary breast cancer between 1981 and 1991. These women have all been followed-up for a minimum of 20 years, and the last date of follow-up was 31st December 2011.

2.2. Statistical methods

2.2.1. Informative censoring

Informative censoring in a cohort of cancer patients is a differential selection process which affects the likelihood of the event of interest being observed. Different strategies have been derived for each data setting and are able to take into account informative censoring when estimating net survival (Appendix A).

Table 1
Description of the two data settings available for the estimation of net survival.

	Setting	Net survival		
			Cause-specific	Relative survival
Biases	Theoretical/Methodological		Informative censoring	
	Data		Misclassification of the cause of death	Non comparability between the cohort and the general population
Solutions	Tackle informative censoring	Concept/Idea	Weight the net survival estimator with the expected mortality	
		Application	Use the cancer data to estimate the expected mortality	Use the expected mortality derived from general population expected mortality
	Check the extent of biases related to the data	Concept/Idea	Sensitivity analyses: Modify the data to check the robustness of the net survival estimate	
		Application	Modify the number of specific death	Modify the expected mortality rates of the general population

The recently proposed Pohar-Perme [6] estimator enables informative censoring to be accounted for in the relative survival data setting, using weights calculated from the expected mortality of each cancer patient according to their individual characteristics. Expected mortality is derived from life tables for the general population from which the cancer patients are drawn and were previously smoothed.

In the cause specific setting, we used a similar strategy to weight the net survival estimator. We derived the weights using the cancer patient data and validated cause of death. We considered that the expected mortality of the cancer patients would be the same as the mortality rate from other causes of death than breast cancer amongst the cancer patients. We fitted a Poisson regression model to the cancer patient data where we considered death from a cause other than breast cancer as the event of interest. We adjusted on age at death and year of death. We used the model to derive expected mortality by age and year. We then used this set of rates to weight the breast-specific mortality hazard, in order to derive net survival estimates.

2.2.2. Potential biases related to the cancer data

In the relative survival setting, the potential bias of interest is related to the comparability between the cancer patients and the general population. An under-estimation of the expected survival would lead in an over-estimation of the net survival. On the contrary, an over-estimation of the expected survival would result in an under-estimation of net survival (Table 2).

In the cause-specific setting, the potential bias of interest is related to the accuracy of the classification of the cause of death. There are two possibilities; the proportion of breast cancer among the deceased patients is either over- or under-reported. If some non-specific deaths are misclassified as breast cancer deaths, the number of deaths from breast cancer is inflated. Net survival is therefore under-estimated. Similarly, net survival is over-estimated when some breast cancer deaths are misclassified as non-breast cancer deaths (Table 2).

2.2.3. Sensitivity analyses

In order to investigate the biases related to each data setting we performed two sensitivity analyses (Table 3).

We defined the baseline situation as the estimation of net survival using the revised and/or validated cause of death in the cause-specific data setting, and the official Geneva life table in the relative survival data setting. We considered that both of these methods derive theoretically unbiased estimates of net survival.

We observed that in the baseline situation (Fig. 1) net survival estimates derived using cause-specific data setting were higher than the relative survival data setting. We therefore concentrated our sensitivity analysis on two of the four potential biases to evaluate how this difference could have arisen (① and ④ Table 2).

In scenario A, we evaluated whether the suitability of the life table in the relative survival data setting might be responsible for the difference observed (④ Table 2). Mortality rates of the general population are only available by age, sex and calendar period in the Geneva canton. Nevertheless, other socio-demographic factors also influence the probability of death for an individual cancer patient. If the life table used in the relative survival setting does not accurately reflect the background risk of death of the cancer patient cohort, biased estimates of survival may result. This can happen, for example, because women with breast cancer tend to be more affluent than the overall population and these affluent women have a lower expected mortality rate than the population of women overall.

Deprivation information about cancer patients is available in three categories in the Geneva Cancer Registry (high, medium and low socio-economic position) but the expected mortality available is not detailed by deprivation. We therefore employed the rate ratios for the first, third and fifth quintiles of deprivation (derived from the England and Wales mortality rates for deprivation quintiles [23]) to build deprivation-specific expected mortality for the three socio-economic groups in the Geneva data. This generated a conservative situation because the differences between these quintiles in the English and Welsh data are likely to be greater than the (unknown) ratios between the three socio-economic groups in Geneva. This resulted in a substantially increased risk of death from other causes for the low socio-economic group and decreased risk for the high socio-economic group (Fig. 2). These deprivation-specific life tables were then used to estimate net survival up to 20 years after diagnosis (scenario A1).

We further evaluated the degree to which expected mortality needed to increase in order to eliminate the difference we observed between the two estimators. We applied the life table for lowest socio-economic group (who have the highest mortality rates) to all

Table 2
Potential biases related to data settings when estimating net survival.

Net survival	Data setting: Cause-specific		Data setting: Relative-survival
Over-estimation	① Real % of BCD (0 — 100) / % of BCD in the data (0 — 100)	②	Under-estimation of the expected survival $E_c < E_p$
Under-estimation	③ Real % of BCD (0 — 100) / % of BCD in the data (0 — 100)	④	Over-estimation of the expected survival $E_c > E_p$

BCD: Breast cancer deaths.
E_p: Expected survival of the general population.
E_c: Expected survival of the cancer patients.

Table 3
Description of the sensitivity analyses performed in order to check the extent of biases related to the data settings.

	Setting	Baseline situation	Scenario					
			A1	A2	B1	B2	B3	B4
Net survival	Cause-specific	Revised and/or validated cause of death	Revised and/or validated cause of death		Percentage of non-specific death reallocated			
					10	15	20	25
	Relative survival	Official Geneva life table	Life table stratified by social class	Life table of the most deprived	Official Geneva life table			

the women in the data and estimated net survival up to 20 years after diagnosis (scenario A2).

We computed the difference between the baseline estimator of net survival using the cause-specific data setting and the relative survival estimators in scenario A1 and A2. Differences were smoothed by running a weighted non-parametric regression on time after diagnosis [24].

In scenario B, we considered misclassification of cause of death as the potential cause of the difference (① Table 2). Since the cause-specific data setting produced the higher of the two estimations, we considered only the situation in which breast cancer deaths had been misclassified as non-specific deaths. In this situation net survival calculated using the cause-specific approach would decrease.

We randomly re-attributed the cause of death variable from non-breast cancer to breast cancer for 10, 15, 20 and 25 per cent of the deceased patients (scenarios B1, B2, B3 and B4, respectively). We iterated this re-attribution 100 times and derived the mean cause-specific net survival up to 20 years for each scenario. The confidence interval was derived using the 95% coverage. The proportion of deaths due to breast cancer among deceased patients varied from 49.7% in the baseline situation to 62.2% in scenario B4 (Table 4).

We computed the difference between the baseline estimator of net survival using the relative survival data setting and the cause-specific estimators for scenarios B1, B2, B3 and B4 respectively. Differences were smoothed by running a weighted non-parametric regression on time after diagnosis [24].

3. Results

The final cohort was comprised of 2489 women diagnosed with an invasive breast cancer between 1981 and 1991 in Geneva, Switzerland.

Fig. 1 shows the baseline situation where net survival estimator using the cause-specific setting was higher than the estimation using relative survival setting for all of the 20 years of follow-up.

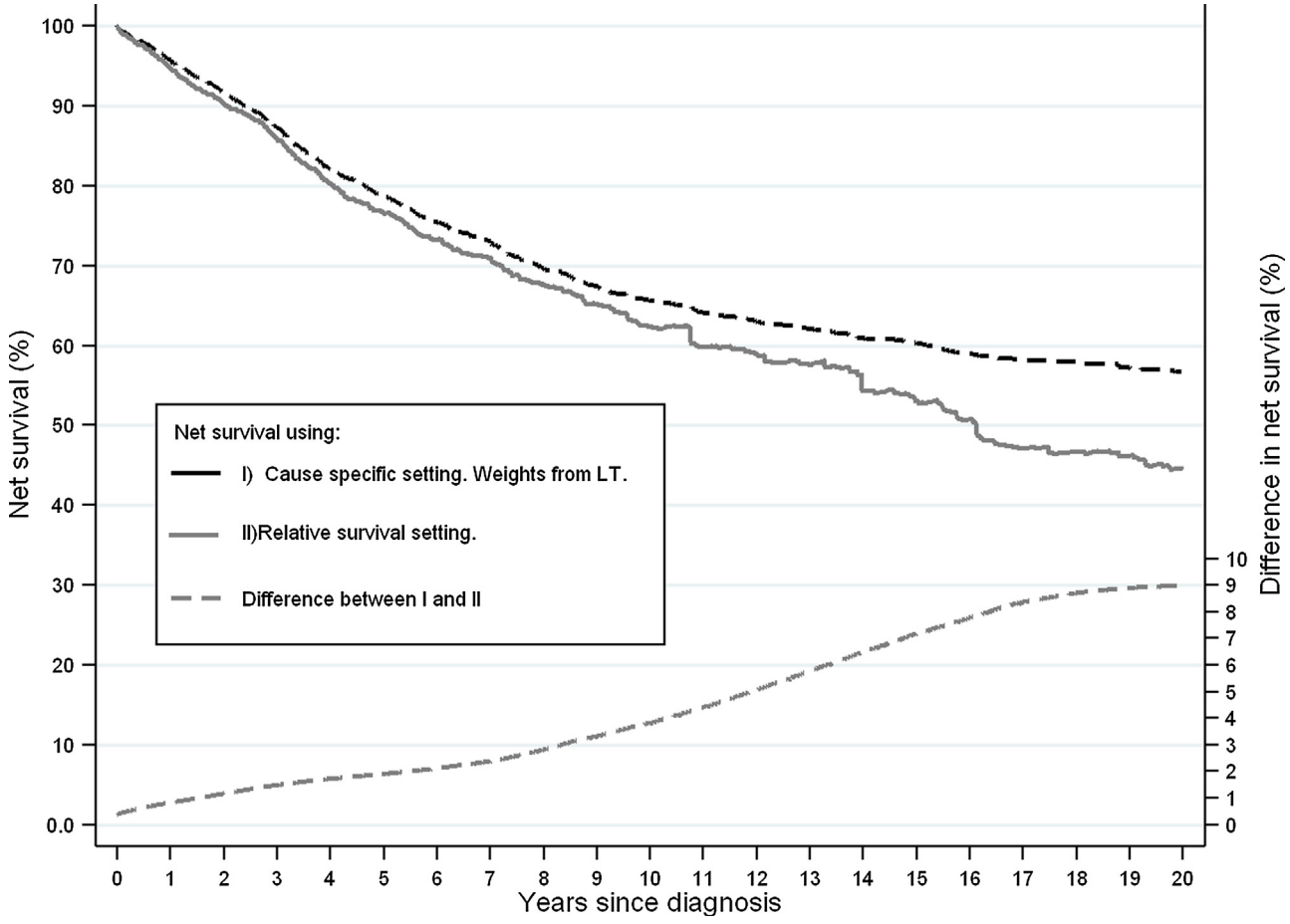


Fig. 1. Net survival estimators in the baseline situation using both cause-specific and relative survival settings.

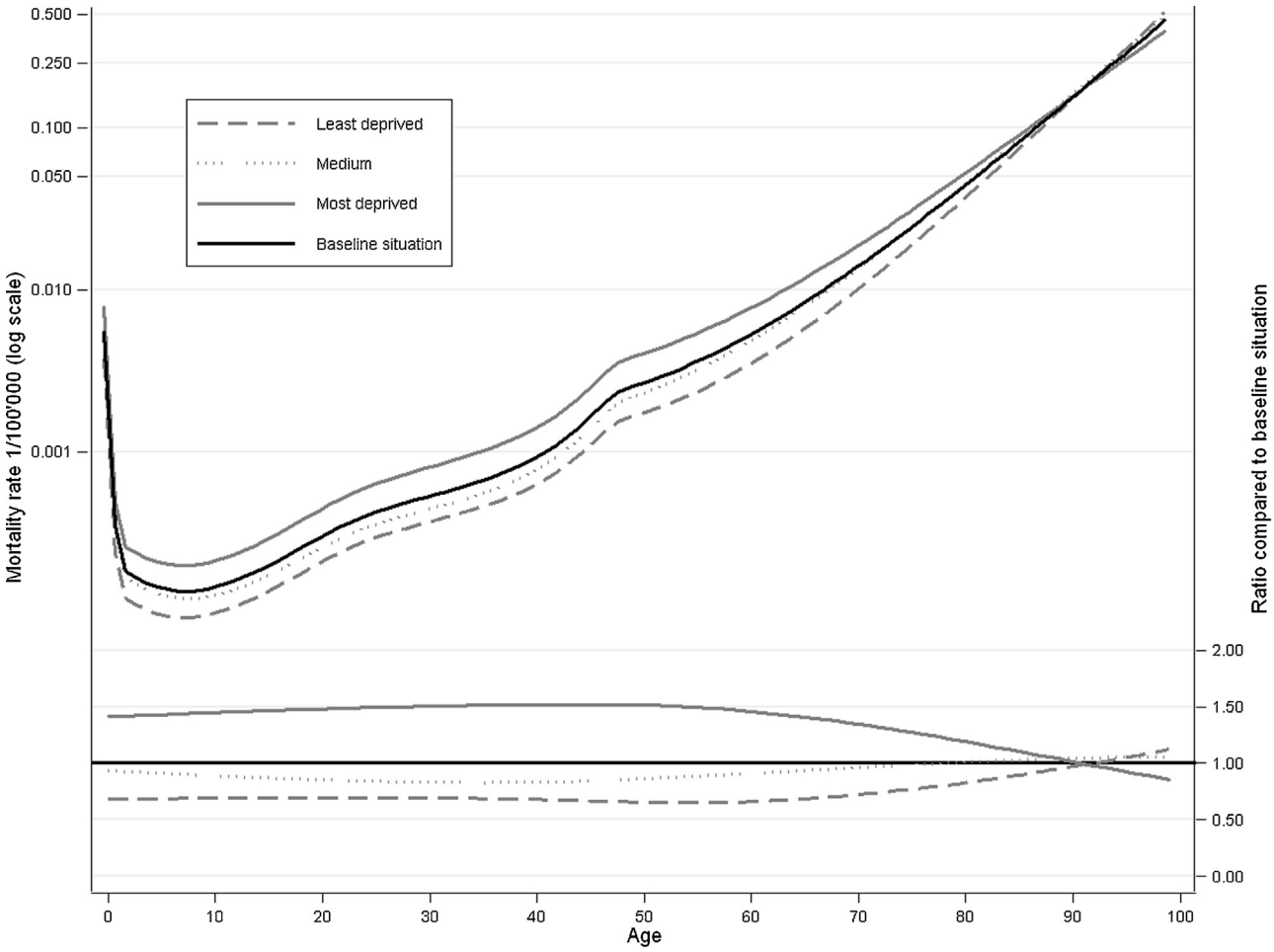


Fig. 2. Geneva general population mortality rates by age for year 1991. Comparison between the baseline situation and scenarios A.

The absolute difference between the two estimators increased with time after diagnosis from 1% at one year to 10.8% at 20 years. It remained less than 3% during the first ten years of follow-up (2.4% at 10 years) and started to increase more dramatically from 13 years onwards (Fig. 1).

In scenario A1, where deprivation-specific life tables were applied, we observed a smaller but still substantial difference between the estimators (Fig. 3). By contrast, the smoothed difference between the two different net survival estimators derived for scenario A2 (use of life tables of the most deprived population) was close to 0 during most of the first ten years after diagnosis (Fig. 3).

Scenarios B1–B4 correspond respectively to the re-allocation of 10, 15, 20 and 25% of deaths from non-breast cancer to breast cancer (Fig. 4). As the proportion of re-allocation increased, the difference between the cause-specific approach and the baseline estimate derived within the relative survival data setting decreased, even turning negative. When 15–20% of the deaths were reallocated, the difference was close to zero. This suggested that with this level of reallocation the cause-specific approach and the relative survival approach used in the baseline situation derived a very similar estimate of net survival up to 10 years after diagnosis. Looking at results after 10 years, the two net survival estimators derived similar estimations when 25% of deaths were reallocated.

4. Discussion

Net survival is the survival that would be observed in a hypothetical world where the only possible underlying cause of death is the disease under study. This study is the first to account for informative censoring in the estimation of net survival in both cause-specific and relative survival data settings, allowing an accurate comparison of two unbiased estimators of net survival. Theoretically, both methods should give the same estimates of net survival. However, both net survival estimates are prone to biases related to the data and their specific assumptions. Differences in the estimates can be attributed to (i) incorrect expected mortality

Table 4
Deaths distribution among female breast cancer patients diagnosed in Geneva between 1981 and 1991 according to scenarios.

	Breast cancer death		Other cause of death		Total number of deaths	
	N[b]	%[b]	*N*[b]	%[b]	*N*[b]	%[b]
% Of deaths reallocated[a]						
Real situation						
0%	844	49.7	856	50.4	1700	100.0
Scenario B1						
10%	930	54.7	770	45.3	1700	100.0
Scenario B2						
15%	972	57.2	728	42.8	1700	100.0
Scenario B3						
20%	1015	59.7	685	40.3	1700	100.0
Scenario B4						
25%	1058	62.2	642	37.8	1700	100.0

[a] Percentage of cases with non-specific cause of death randomly recoded as breast cancer.
[b] The numbers (*N*) and percentages (%) given are an average of the 100 iterations for scenarios B1, B2, B3 and B4.

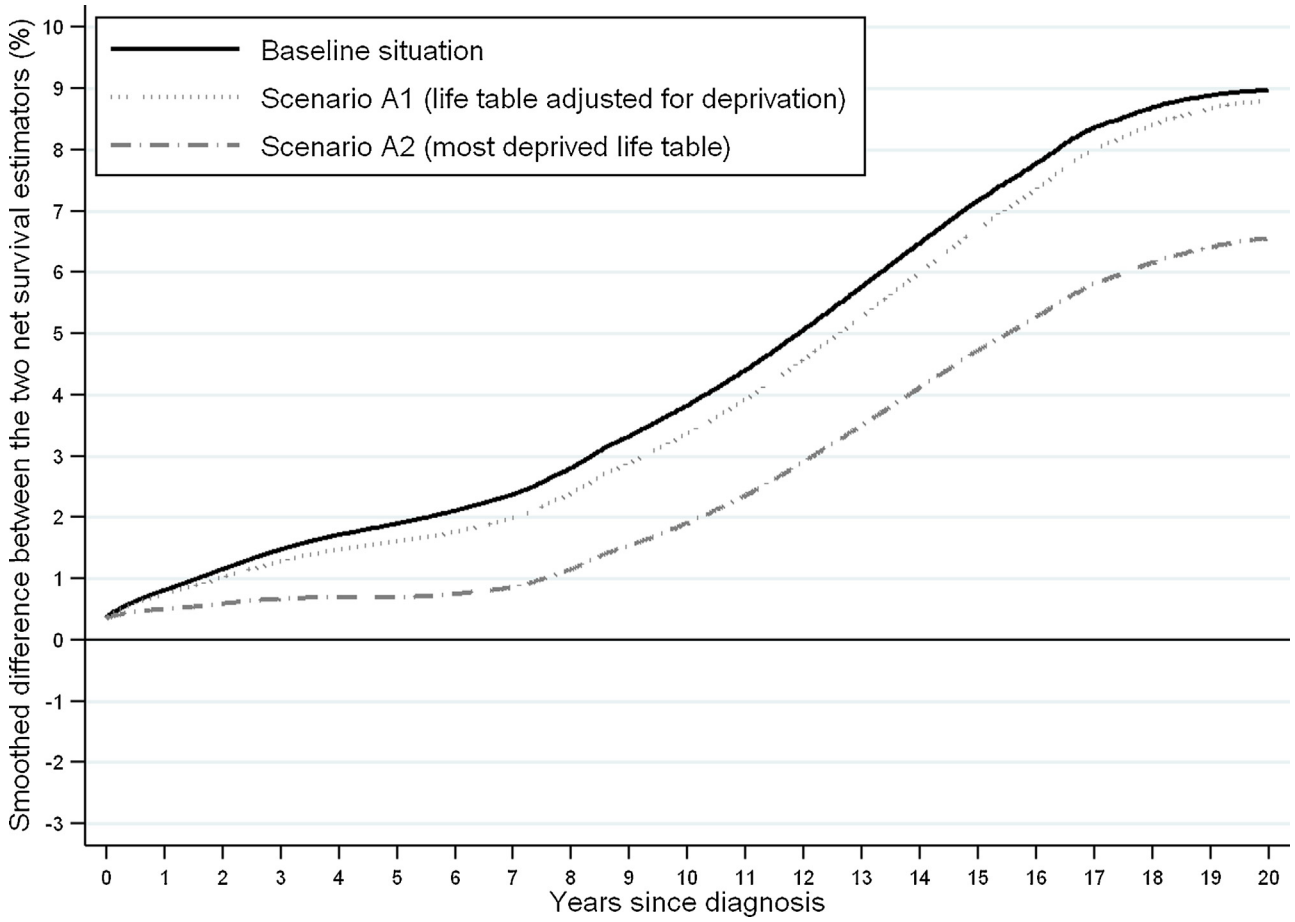


Fig. 3. Smoothed differences between the net survival estimators when informative censoring is taken into account: Cause-specific setting in the baseline situation vs. Relative survival setting in baseline situation, scenario A1 and scenario A2.

due to inadequate life tables in the relative survival data setting, or (ii) errors in the cause of death for some patients in the cause-specific data setting. We have evaluated these two possibilities using data on breast cancer patients whose cause of death has been independently validated, and who have been followed for 20 years after their diagnosis.

In the cause-specific setting, weights, estimated using the mortality hazard specific to the other causes of death from the cancer registry data, were applied to tackle informative censoring and estimate theoretically unbiased net survival. Although these internal weights were derived from a model, they may have been unstable due to small numbers of deaths in this fairly small breast cancer population. We thus also derived the weights using the expected mortality from the general population life tables (therefore similar as those used with the relative survival setting). Both weighting approaches gave very similar results, confirming the strength of the cause-specific setting estimator. We assumed that the weighted net survival estimator in the cause-specific setting was theoretically unbiased and equivalent to the Pohar Perme approach. However, simulation-based work is needed to assess its performance.

We observed that the cause-specific approach gave higher estimates of net survival compared to the relative survival approach (Fig. 1). Moreover, the absolute difference between the two estimators increased with time since diagnosis up to 2.5% at 10 years after diagnosis and over 10% at 20 years. This enabled us to consider only two (① and ④, Table 2) out of the four potential biases related to the data setting. Indeed, we did not evaluate option ② or ③. Option ② considers the situation in which net survival is over-estimated because of under-estimated expected survival. This situation is unlikely insofar as cancer survivors are often prone comorbidities and are therefore no less likely to die than the general population, even if at longer term, the situation may be reversed, with the more robust patients being selected [25,26]. Option ③ describes the situation in which net survival is under-estimated because deaths not due to breast cancer are mistakenly classified as breast cancer deaths. It is however more likely that the true number of breast cancer deaths is under-estimated because deaths caused indirectly by breast cancer may be misclassified.

We first estimated net survival in the relative survival setting using different life tables to evaluate whether non-comparability between the general population and the cohort of patient under study compromises the estimation of net survival. In our study, we stratified the life table on deprivation (scenario A1) and the results showed that the net survival estimation was not substantially altered by this. In scenario A2, we noticed that a very large increase in the mortality rate was required (30% for a 50 year old woman) before the net survival estimated in the relative survival setting reached the net survival estimated in the cause-specific setting during the first 8 years after diagnosis. Such changes seem quite unreasonable at a population level. Net survival in the relative survival setting appeared therefore to be robust to inaccuracies in the underlying mortality rate.

The second sensitivity analysis showed that the level of misclassification could be relatively small to observe a

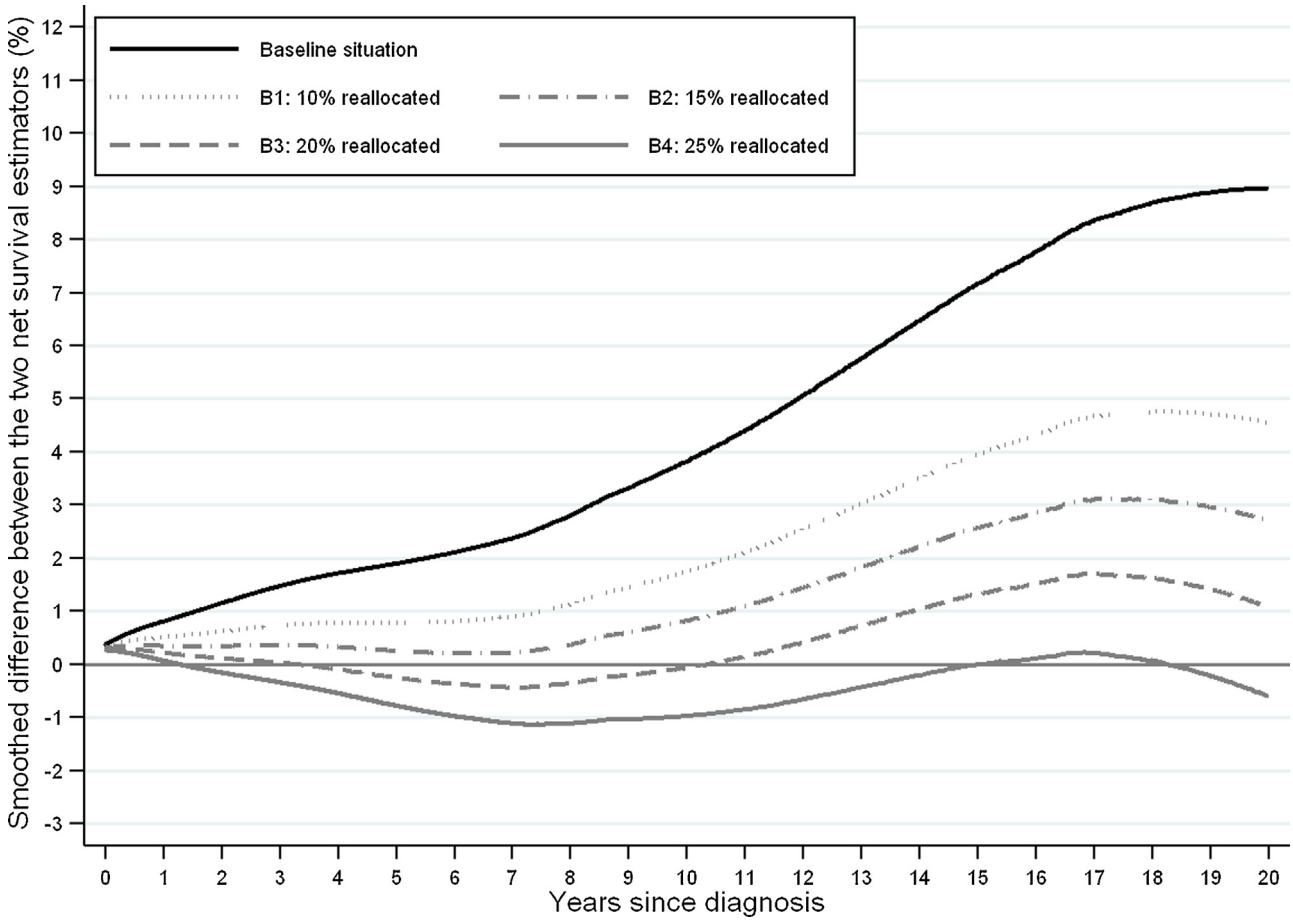


Fig. 4. Smoothed differences between the net survival estimators when informative censoring is taken into account: Cause-specific setting in the baseline situation, scenarios B1, B2, B3 and B4 vs. Relative survival setting in the baseline situation.

large change in net survival. Indeed, recoding only 10% of the non-specific deaths led to a large decrease of the net survival using the cause-specific settings. The estimator of net survival derived with cause-specific method proved therefore to be relatively sensitive to the allocation of the cause of death. Recoding 15–20% of deaths from other causes to breast cancer resulted in the convergence of the net survival estimates in both settings. We have previously shown that a review of clinical files resulted in cause of death being revised for 9 per cent of women with breast cancer in the Geneva Cancer Registry data [22]. In the current study, 20% of the non-specific deaths represented 171 cases out of the 1700 deceased patients. As such, survival estimates are likely to be biased by the misclassification of a relatively small number of deaths that are indirectly related to breast cancer (for instance; side effects of treatment or suicide). The fact that the proportion of reallocation required to reduce the survival difference to zero increased after 10 years after diagnosis (from 10% to 15%) lends weight to this argument insofar that allocating breast cancer as an underlying cause of death is less probable with increasing time since diagnosis.

Taken together, these results suggest that survival derived with the cause-specific approach provides a very sensitive estimator, likely to be an overestimate of true net survival. On the contrary, survival derived with the relative survival approach is likely to be closer to the actual net survival of the patient cohort insofar as it is very robust to changes to the expected rate of death. This is especially true with increasing time after diagnosis.

Our study has considered only women with breast cancer. Breast cancer patients are not, however, representative of patients with cancer at different localisations. The proportions of specific deaths and age have a large impact in biases related to net survival. Future work will test the repeatability of our analyses on other cancer sites by different age groups. Preliminary results on cancers of colon-rectum, lung and on melanoma suggested results consistent with those provided by this study.

A dramatic increase in long-term survivors has been observed over the last few decades as a result of screening programs, more precise diagnostic tools and developments in treatment protocols [27]. In the future a particular interest will be given to long-term net survival estimation, especially among younger patients.

Our results suggest that, when analysing routinely collected population-based data, the relative survival setting is likely to derive more accurate estimates of net survival, and that the cause specific setting is vulnerable to misclassification bias, particularly in the long-term. The relative survival setting is therefore highly recommended when estimating net survival with population-based data.

Funding

This work was supported by the Swiss Cancer League [BIL KFS-3274-08-2013].

Conflict of interest

The authors have declared no conflict of interest.

Authorship contribution

All authors contributed to the manuscript. RS conducted the analysis and the writing under the supervision of LW and BR. BR, LW and AB reviewed the paper and made final corrections. All authors read and approved the final version of the manuscript.

Acknowledgements

This collaborative work was supported by the Geneva Cancer Registry and the Cancer Survival Group at London School of Hygiene and Tropical Medicine. We should like to thank Massimo Usel, Elisabetta Rapiti, Hyma Schubert, Christine Bouchardy and all collaborators at the Geneva Cancer Registry as well as Michel Coleman and all collaborators at the London School of Hygiene and Tropical Medicine for their support and encouragements. This work, presented at the GRELL Meeting 2014 in Geneva, was awarded the 'Enrico Anglesio' Prize, offered by the 'Anglesio Moroni Foundation', Turin, Italy.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.canep.2015.04.001.

References

[1] Berkson J, Gage RP. Calculation of survival rates for cancer; 1952.
[2] Ederer F, Cutler SJ. The relative survival rate: a statistical methodology. Natl Cancer Inst Monogr 1961;6:101–21.
[3] Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: the importance of allowing for informative censoring. Stat Med 2012;31(8):775–86.
[4] Robins J. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. Proc Biopharm Sect Am Stat Assoc 1993;24–33.
[5] Satten G. Estimating the marginal survival function in the presence of time dependent covariates. Stat Probab Lett 2001;54(4):397–403.
[6] Perme MP, Stare J, Estève J. On estimation in relative survival. Biometrics 2012;68(1):113–20.
[7] Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. Am J Public Health 1981;71(3):242–50.
[8] Hoel DG, Ron E, Carter R, Mabuchi K. Influence of death certificate errors on cancer mortality trends. J Natl Cancer Inst 1993;85(13):1063–8.
[9] Messite J, Stellman SD. Accuracy of death certificate completion: the need for formalized physician training. JAMA 1996;275(10):794–6.
[10] Maudsley G, Williams EM. 'Inaccuracy' in death certification – where are we now? J Public Health Med 1996;18(1):59–66.
[11] Lee PN. Comparison of autopsy, clinical and death certificate diagnosis with particular reference to lung cancer. A review of the published data. APMIS Suppl 1994;45:1–42.
[12] Lloyd-Jones DM, Martin DO, Larson MG, Levy D. Accuracy of death certificates for coding coronary heart disease as the cause of death. Ann Intern Med 1998;129(12):1020–6.
[13] Newschaffer CJ, Otani K, Mcdonald MK, Penberthy LT. Causes of death in elderly prostate cancer patients and in a comparison nonprostate cancer cohort. J Natl Cancer Inst 2000;92(8):613–21.
[14] Goldoni CA, Bonora K, Ciatto S, Giovannetti L, Patriarca S, Sapino A, et al. Misclassification of breast cancer as cause of death in a service screening area. Cancer Causes Control 2009;20(5):533–8.
[15] Crocetti E, De Lisi V, Gafà L, Sechi O, Mangone L. Net survival: comparison between relative and cause-specific survival estimates. Epidemiol Prev 2001;25(3 Suppl.):32–6.
[16] Stelzner S, Hellmich G, Koch R, Witzigmann H. Exactitude of relative survival compared with cause-specific survival and competing risk estimations based on a clinical database of patients with colorectal carcinoma. Dis Colon Rectum 2009;52(7):1264–71.
[17] Sarfati D, Blakely T, Pearce N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. Int J Epidemiol 2010;39(2):598–610.
[18] Baade PD, Fritschi L, Eakin EG. Non-cancer mortality among people diagnosed with cancer (Australia). Cancer Causes Control 2006;17(3):287–97.
[19] Ezzati M, Lopez AD. Estimates of global mortality attributable to smoking in 2000. Lancet 2003;362(9387):847–52.
[20] Zahl PH, Tretli S. Long-term survival of breast cancer in Norway by age and clinical stage. Stat Med 1997;16(13):1435–49.
[21] Ellis L, Coleman MP, Rachet B. The impact of life tables adjusted for smoking on the socioeconomic difference in net survival for laryngeal and lung cancer. Br J Cancer 2014;111:195–202.
[22] Schaffar R, Rapiti E, Rachet B, Woods L. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva Cancer Registry. BMC Cancer 2013;13(2):609.
[23] Woods LM, Rachet 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(2):115–20.
[24] Cleveland W. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 1979;74:829–36.
[25] Zahl PH. Frailty modelling for the excess hazard. Stat Med 1997;16(14):1573–85.
[26] Riihimäki M, Thomsen H, Brandt A, Sundquist J, Hemminki K. Death causes in breast cancer patients. Ann Oncol 2012;(23):604–10.
[27] Soerjomataram I, Louwman MWJ, Ribot JG, Roukema JA, Coebergh JWW. An overview of prognostic factors for long-term survivors of breast cancer. Breast Cancer Res Treat 2008;107(3):309–30.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Robin Schaffar
Principal Supervisor	Dr Laura Woods
Thesis Title	Long-term net survival among women diagnosed with breast cancer: accuracy of its estimation and evaluation of its determinants

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	European Journal of Cancer		
When was the work published?	2017		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. I planned and carried out the literature review and data analysis. I prepared all drafts of the paper. The co-authors provided input and feedback on the content data analysis and on the paper drafts prepared by me.

Student Signature: [redacted] **Date:** 09/03/2018

Supervisor Signature: [redacted] **Date:** 09/03/2018

European Journal of Cancer 72 (2017) 78–83

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Estimation of net survival for cancer patients: Relative survival setting more robust to some assumption violations than cause-specific setting, a sensitivity analysis on empirical data

Robin Schaffar [a,b,*], Bernard Rachet [b], Aurélien Belot [b], Laura M. Woods [b]

[a] *Geneva Cancer Registry, Global Health Institute, University of Geneva, Geneva, Switzerland*
[b] *Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom*

Received 15 July 2016; received in revised form 27 October 2016; accepted 15 November 2016
Available online 24 December 2016

KEYWORDS
Net survival;
Cause-specific;
Relative survival;
Informative censoring

Abstract Net survival is the survival that would be observed if the only possible underlying cause of death was the disease under study. It can be estimated with either cause-specific or relative survival data settings, if the informative censoring is properly considered. However, net survival estimators are prone to specific biases related to the data setting itself. We examined which data setting was the most robust against violation of key assumptions (erroneous cause of death and inappropriate life tables).

We identified 4285 women in the Geneva Cancer Registry, diagnosed with breast, colorectal, lung cancer and melanoma between 1981 and 1991 and estimated net survival up to 20 years using cause-specific and relative survival settings. We used weights to tackle informative censoring in both settings and performed sensitivity analyses to evaluate the impact of misclassification of cause of death in the cause-specific setting or of using inappropriate life tables on net survival estimates in the relative survival setting.

For all the four cancers, net survival was highest when using the cause-specific setting and the absolute difference between the two estimators increased with time since diagnosis. The sensitivity analysis showed that (i) the use of different life tables did not compromise net

* *Corresponding author*: Geneva Cancer Registry, Global Health Institute, University of Geneva, 55 Boulevard de la Cluse, 1205 Geneva, Switzerland.
E-mail address: robin.schaffar@unige.ch (R. Schaffar).

http://dx.doi.org/10.1016/j.ejca.2016.11.019

survival estimation in the relative survival setting, whereas (ii) a small level of misclassification for the cause of death led to a large change in the net survival estimate in the cause-specific setting.

The relative survival setting was more robust to the above assumptions violations and is therefore recommended for estimation of net survival.

1. Introduction

Net survival measures the survival that would be observed if the only possible cause of death was the disease of interest [1]. It is the most defensible method of estimating survival from cancer. Two main settings have been described for its estimation: the relative survival setting and the cause-specific setting. The latter requires information on underlying cause of death so that deaths due to causes other than the cancer of interest can be censored. Such information is not needed in the relative survival setting. Here, the overall survival of the cancer patients is compared with the survival they would have experienced if they had had the mortality of the general population from which they were drawn [2].

In both settings, net survival estimation is susceptible to bias due to informative censoring. Informative censoring occurs when patients are removed from the risk set (censored) under a non-random way: these patients would experience a different mortality hazard compared with those that remain in the risk set [3]. In the cause-specific setting, when the interest is in estimating the cancer-specific mortality hazard, patients who died due to other cause are censored (and so removed from the risk set). It means that patients with higher risk of dying from causes other than cancer (for example, elderly compare to young patients) are more likely to be removed from the risk set. However, because age is also an important prognostic factor for cancer, censoring these patients is informative for the cancer-survival estimation. In the relative survival setting, this mechanism of informative censoring is less easy to conceptualise (because the cause of death is unknown and/or not used); any variable with an effect on both cancer-specific and other cause mortality hazards induces informative censoring. Demographic variables which define the life tables may lead to informative censoring and need to be accounted for. A new estimator has been described by Pohar-Perme which is able to take account of this bias within the relative survival setting [4] and its performances have been assessed in an extensive simulation study [5]. We have recently proposed a similar strategy for the estimation of net survival in the cause-specific setting [6].

If informative censoring is accounted for, estimates of net survival derived in each of these settings are theoretically unbiased. However, biases relating to the data setting itself may still occur. In the relative survival setting, bias can originate from the non-comparability between the cohort and the general population from which rates of expected mortality are drawn, due to unmeasured variable(s) affecting both expected and excess hazard rates (this latter being the rate from which the net survival is derived). In the cause-specific setting, bias can arise from the misclassification of the underlying cause of death. Our previous analyses of patients diagnosed with breast cancer in Geneva showed that the estimation of net survival using the cause-specific setting was very sensitive to the codification of underlying cause of death, but, in contrast, the relative survival setting was robust to non-comparability in the estimation of background mortality [6].

Breast cancer may, however, represent a special case. Survival among breast cancer patients is high, but deaths directly caused by the original cancer still occur into the second and third decades following diagnosis: a pattern of excess mortality which is seen for very few other anatomic sites. As such, our previous conclusion may not hold for every cancer type. Here, we extend our analysis of breast cancer patients to patients diagnosed with cancers of three other anatomic sites (according to the international classification of disease, 10th version, ICD-10)to establish whether the same conclusions hold for other malignancies.

2. Material and methods

The Geneva Cancer Registry records underlying cause of death for all cancer patients. More unusual, the registry also validates the accuracy of this variable by reviewing all clinical information available for each patient. The overall agreement between the variables (revised cause of death versus cause of death based on death certificates) was high. However, several subgroups presented a lower concordance, suggesting differences in calendar time and less attention given to older patients and more advanced diseases [7]. This context thus represents a unique opportunity to compare relative survival and cause-specific settings when estimating net survival, because the registry holds more accurate information on the underlying cause of death.

Table 1
Deaths distribution by age groups, years after diagnosis and cancer sites for women diagnosed with an invasive tumour between 1981 and 1991.

Year of follow up		Colon rectum (N = 996, mean age: 72.1)												Lung (N = 500, mean age: 67.6)									
		0–1		2–5		6–10		11–15		16–20		All		0–1		2–5		6–10		11–15		16–20		All	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
All patients	Deceased	345	40	264	30	95	11	78	9	45	5	**868**	**100**	320	66	124	26	21	4	5	1	9	2	**483**	**100**
	Others	*53*	*15*	*85*	*32*	*65*	*68*	*70*	*90*	*44*	*98*	***357***	***41***	*25*	*8*	*20*	*16*	*8*	*38*	*3*	*60*	*6*	*67*	***64***	***13***
	Cancer-specific	*292*	*85*	*179*	*68*	*30*	*32*	*8*	*10*	*1*	*2*	***511***	***59***	*295*	*92*	*104*	*84*	*13*	*62*	*2*	*40*	*3*	*33*	***419***	***87***
<50	Deceased	9	31	11	38	5	17	1	3	–	–	**29**	**100**	29	60	12	25	3	6	0	0	1	2	**48**	**100**
	Others	*3*	*33*	*2*	*18*	*0*	*0*	*1*	*100*	-	-	***9***	***31***	*1*	*3*	*2*	*17*	*0*	*0*	*0*	*0*	*2*	*67*	***5***	***10***
	Cancer-specific	*6*	*67*	*9*	*82*	*5*	*100*	*0*	*0*	-	-	***20***	***69***	*28*	*97*	*10*	*83*	*3*	*100*	*1*	*100*	*1*	*33*	***43***	***90***
50–69	Deceased	81	37	70	32	13	6	19	9	11	5	**220**	**100**	117	61	55	29	9	5	3	2	8	4	**193**	**100**
	Others	*10*	*12*	*18*	*26*	*9*	*69*	*18*	*95*	*10*	*91*	***90***	***41***	*2*	*2*	*11*	*20*	*3*	*33*	*2*	*67*	*6*	*75*	***24***	***12***
	Cancer-specific	*71*	*88*	*52*	*74*	*4*	*31*	*1*	*5*	*1*	*9*	***130***	***59***	*115*	*98*	*44*	*80*	*6*	*67*	*1*	*33*	*2*	*25*	***169***	***88***
70+	Deceased	255	41	183	30	77	12	59	10	33	5	**619**	**100**	174	72	57	24	9	4	2	1	–	–	**242**	**100**
	Others	*40*	*16*	*65*	*36*	*56*	*73*	*52*	*88*	*33*	*100*	***258***	***42***	*22*	*13*	*7*	*12*	*5*	*56*	*1*	*50*	-	-	***35***	***15***
	Cancer-specific	*215*	*84*	*118*	*65*	*21*	*27*	*7*	*12*	*0*	*0*	***361***	***58***	*152*	*87*	*50*	*88*	*4*	*44*	*1*	*50*	-	-	***207***	***86***

Year of follow up		Melanoma (N = 300, mean age: 53.1)												Breast (N = 2489, mean age: 62.1)									
		0–1		2–5		6–10		11–15		16–20		All		0–1		2–5		6–10		11–15		16–20		All	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
All patients	Deceased	10	7	37	27	30	22	20	15	19	14	**137**	**100**	181	11	574	34	417	25	228	13	176	10	**1700**	**100**
	Others	*4*	*40*	*20*	*54*	*18*	*60*	*17*	*85*	*17*	*90*	***97***	***71***	*76*	*42*	*205*	*36*	*189*	*45*	*150*	*66*	*130*	*74*	***856***	***50***
	Cancer-specific	*6*	*60*	*17*	*46*	*12*	*40*	*3*	*15*	*2*	*11*	***40***	***29***	*105*	*58*	*369*	*64*	*228*	*55*	*78*	*34*	*46*	*26*	***844***	***50***
<50	Deceased	1	5	8	36	8	36	0	0	1	5	**22**	**100**	8	3	78	32	70	29	34	14	32	13	**245**	**100**
	Others	*0*	*0*	*1*	*13*	*0*	*0*	*0*	*0*	*1*	*100*	***5***	***23***	*0*	*0*	*12*	*15*	*17*	*24*	*11*	*32*	*12*	*38*	***68***	***28***
	Cancer-specific	*1*	*100*	*7*	*88*	*8*	*100*	*1*	*100*	*0*	*0*	***17***	***77***	*8*	*100*	*66*	*85*	*53*	*76*	*23*	*68*	*20*	*63*	***177***	***72***
50–69	Deceased	4	8	11	22	6	12	8	16	7	14	**49**	**100**	56	9	205	33	133	21	75	12	77	12	**624**	**100**
	Others	*1*	*25*	*5*	*46*	*4*	*67*	*6*	*75*	*6*	*86*	***35***	***71***	*13*	*23*	*50*	*24*	*31*	*23*	*41*	*55*	*56*	*73*	***260***	***42***
	Cancer-specific	*3*	*75*	*6*	*55*	*2*	*33*	*2*	*25*	*1*	*14*	***14***	***29***	*43*	*77*	*155*	*76*	*102*	*77*	*34*	*45*	*21*	*27*	***364***	***58***
70+	Deceased	5	8	18	27	16	24	12	18	11	17	**66**	**100**	117	14	291	35	214	26	119	14	67	8	**831**	**100**
	Others	*3*	*60*	*14*	*78*	*14*	*88*	*11*	*92*	*11*	*100*	***57***	***86***	*63*	*54*	*143*	*49*	*141*	*66*	*98*	*82*	*62*	*93*	***528***	***64***
	Cancer-specific	*2*	*40*	*4*	*22*	*2*	*13*	*1*	*8*	*0*	*0*	***9***	***14***	*54*	*46*	*148*	*51*	*73*	*34*	*21*	*18*	*5*	*8*	***303***	***37***

Italic figures detail the number/proportions of deaths by cause.
Bold figures represent the number/proportions of deaths for the entire period of follow-up.

We selected women diagnosed between 1981 and 1991 at ages 15 to 99 with invasive colorectal (C18–20), lung (C34), melanoma (C44.1) or breast (C50) cancer. These malignancies afforded us tumours with a wide range of aggressiveness as well as very different incident age distributions. All patients were followed up to the end of 2012.

Our approach has been described previously [6]. Briefly, we used the Pohar-Perme estimator in the relative survival setting and our own derived estimator for the cause-specific setting to estimate net survival. Both estimators take into account informative censoring, that is the fact that the number of patients we observed to be at risk is smaller than the number of patients that would be at risk in the hypothetical world, were people could die of the cancer of interest only. Because the same is true for the number of deaths as well weights are used to correct the net survival estimates for this bias [8,9]. We define these estimates as the 'baseline situation'.

We then examined two sets of scenarios to evaluate the extent of biases arising from the data setting. The aim of scenarios A1 and A2 was to evaluate the impact of the life tables on net survival estimation within the relative survival setting. In Geneva, general population mortality rates are available by year of age and calendar year. However, other socio-demographic variables are known to have a strong influence on the probability of death. In scenario A1, we consider a simulated stratification of the expected age-, sex- and period-specific mortality rates by deprivation. In scenario A2, we artificially increase the expected mortality of all the patients, well above what would ordinarily be expected, by attributing the mortality of the most deprived patients to the whole cohort. We computed the difference between the baseline estimator (using the cause-specific setting) and the net survival estimates derived under scenarios A1 and A2. Differences were smoothed with a weighted non-parametric regression on time since diagnosis [10].

Scenario B aimed to evaluate the impact of misclassifying the cause of death on net survival estimation in the cause-specific setting. Here, we randomly reattributed non-cancer deaths to cancer deaths for 10, 15, 20 and 25% of the deceased patients (scenarios B1, B2, B3 and B4, respectively). This was performed 100 times to derive a mean cause-specific net survival for each scenario. We derived the difference between the baseline estimator (using the relative survival setting) and the

cause-specific estimates in scenarios B1, B2, B3 and B4. Differences were smoothed with a weighted non-parametric regression on time since diagnosis [10].

3. Results

The final cohort was composed of 996 women diagnosed with colorectal cancer, 500 women diagnosed with lung cancer, 300 women diagnosed with melanoma and 2489 women diagnosed with breast cancer.

Table 1 describes the age distribution and aggressiveness of each disease. Patients diagnosed with colorectal, lung, melanoma and breast cancer presented a mean age of 72.1, 67.6, 53.1 and 62.1 years, respectively. For colorectal cancers, 87% of the patients died, 511 of their cancer (59%). Among women diagnosed with lung cancer, 97% died. There were 483 deaths, 419 from lung cancer (87%). Among patients with melanoma, 46% died, 40 due to melanoma (29%). For breast cancer, 1700 patients died (68%) 844 from their cancer (50%).

Baseline estimators of net survival are presented in Fig. 1 for each cancer site. We observed, consistent with our previous analyses, that net survival estimates using the cause-specific setting are higher than the estimates using the relative survival setting for every localisation. The absolute difference between the two estimators increased with time since diagnosis for all four cancers. For colorectal cancer, the difference widened from almost 2% at one year to over 7% at 20 years. For lung cancer, the difference increased sharply within the first two years after diagnosis (3% at two years) and moderately afterwards. There was no detectable difference during the first three years after diagnosis for melanoma, but it subsequently increased to more than 8% at 20 years after diagnosis. For breast cancer, the absolute difference between the two estimators increased with time since diagnosis from 1% at one year to 11% at 20 years.

Where deprivation-specific life tables were used (scenario A1), the difference between both net survival estimators was fairly constant across all four cancers (Fig. 2). When we used the life tables of the most deprived population (scenario A2), we still observed a small but substantial difference for all cancer sites.

By contrast, increasing the proportion of deaths classified as being due to cancer led to a decreasing difference between the cause-specific estimate and the

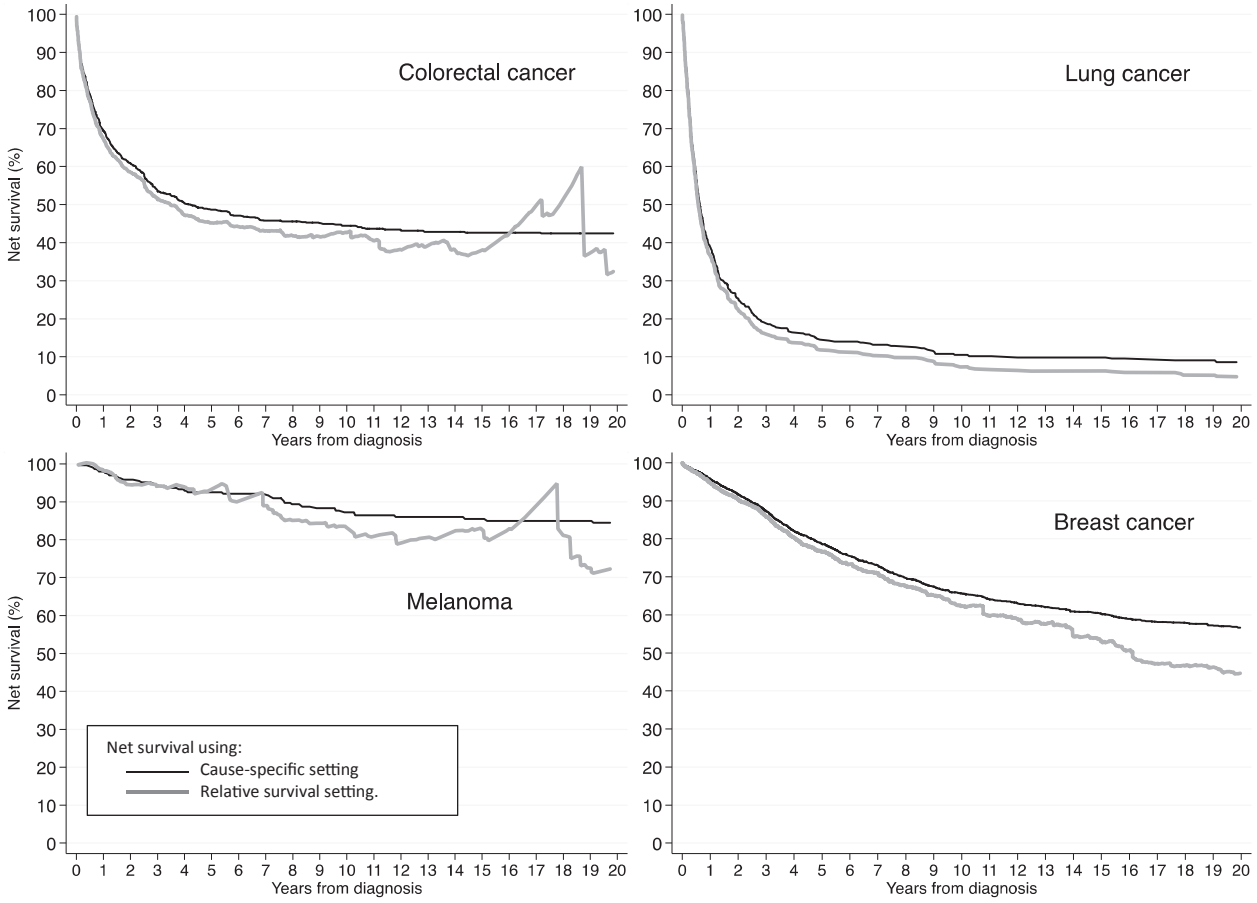


Fig. 1. Estimation of net survival for the four localisations using both cause-specific and relative survival setting. Geneva Cancer Registry, 1981–1991.

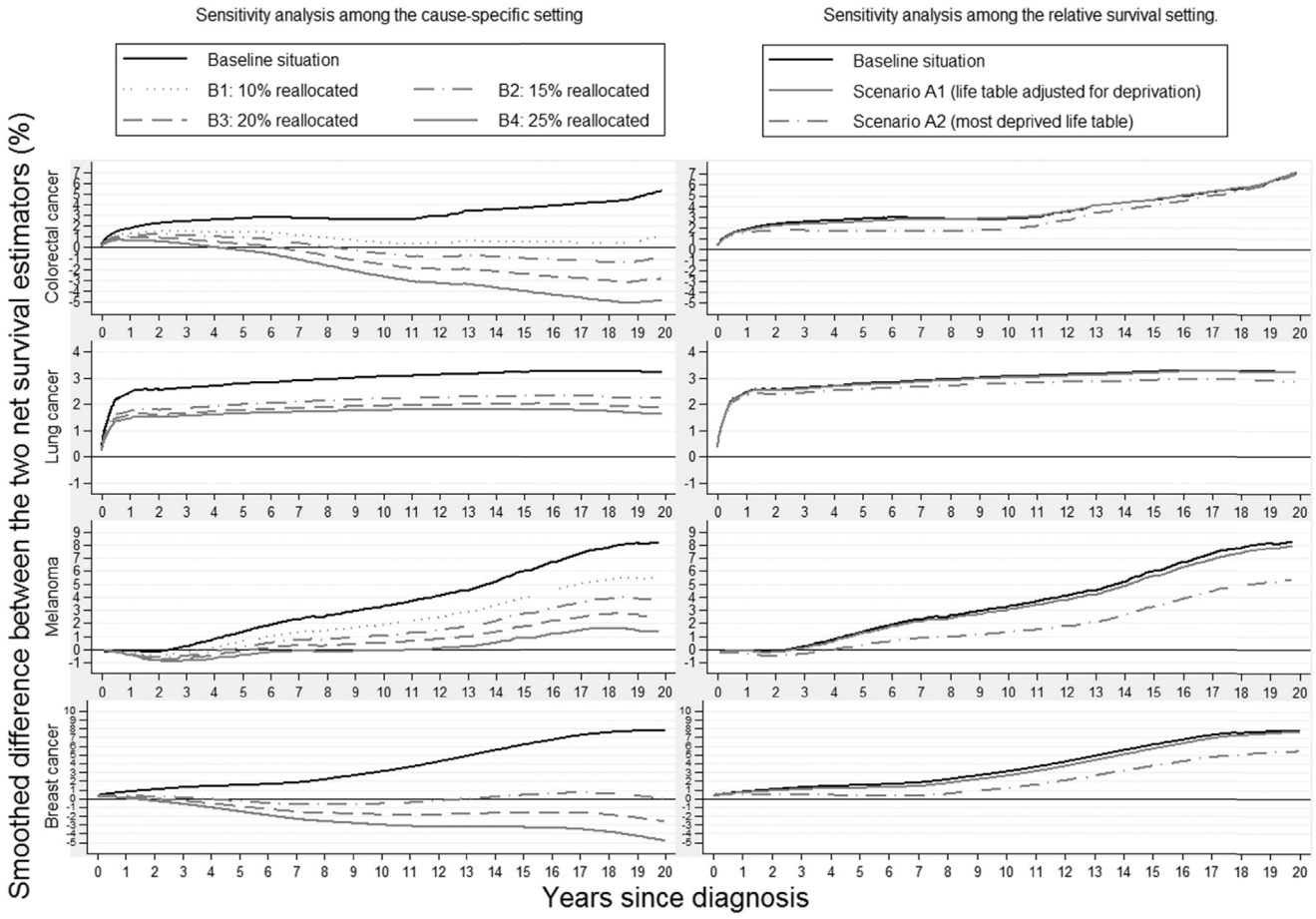


Fig. 2. Sensitivity analysis among both cause-specific and relative survival setting. Geneva Cancer Registry, 1981–1991.

baseline estimate (derived within the relative survival setting) for all anatomic sites, even turning negative for colon cancer and melanoma (Scenarios B1–B4). For colorectal cancer, the effect was dependent on time since diagnosis: early in follow up the difference was eliminated only with more than 25% of deaths reallocated. However, after 5 years, the re-allocation of 10–15% resulted in no difference in the two estimators. In contrast, for lung cancer, 25% re-allocation did not eliminate the difference between the two net survival estimators. From 6 years after diagnosis, the difference was almost eliminated for melanoma if 25% of deaths were reallocated. For breast cancer, the difference decreased as the proportion of re-allocation increased, even turning negative. When 15–20% of the deaths were reallocated, the difference was close to zero.

4. Discussion

This study has evaluated whether our previous conclusions relating to the nature and size of modifications to the data in each setting are similar for breast cancer [6] and other cancers with very different patterns of incidence and excess mortality. We account for informative censoring in all analyses, allowing an accurate comparison of two unbiased net survival estimators. Our analyses are therefore not comparable with previous studies comparing cause-specific survival and relative survival [11,12]. Indeed, the estimators used for these studies were not estimating net survival as they did not account for informative censoring [4].

Differences between the two net survival estimators varied with time since diagnosis and with anatomic site. However, the cause-specific approach generally resulted in higher estimates of net survival. The first sensitivity analysis (scenarios A1 and A2) showed that the use of different life tables did not compromise net survival estimation in the relative survival setting. Even with a very large modification in the expected mortality rate (by the application of mortality rates for the most deprived population to all patients), estimates of net survival were fairly stable. Net survival estimation in the relative survival setting appeared, therefore, to be relatively robust to non-comparability of the underlying mortality rates to the patient population, irrespective of the anatomic site. By contrast, the second sensitivity analyses showed a greater impact in net survival estimates within the cause-specific setting: a relatively small

level of misclassification for the underlying cause of death led to a large change in the net survival estimate. This was true for all cancer sites.

After longer periods of follow up, the Pohar-Perme estimator tended to produce erratic results when the number of deaths was small. Net survival can increase within the relative survival setting because the observed mortality of the cancer patient group can be lower than their expected mortality. With increasing time, the few remaining patients need to represent more and more of their counterparts; such patients, especially among elderly, are more likely to survive better than the general population, resulting in overall hazard lower than expected hazard. The excess hazard therefore becomes negative and the survival function increases. The erratic curves in the context of net survival derived with the relative survival setting therefore originates from the fact that we are asking questions about the hypothetical world that are not supported by sufficient information in the real world. We estimated net survival at 20 years for comparative purpose but being interested in a '25-year net survival of a 90-year old patient' implies asking what would happen in 25 years to a patient who is 90 at the time of diagnosis if they could not die from other reasons than cancer. Such a question of course makes no sense. Therefore, the length of analytical follow-up time should be restricted so that the population survival probability for all the patients in the cohort is large enough. Another difficulty in long-term (net) survival is due to the increasing probability of multiple tumours. In our study, for patients having several tumours with the same ICDO-code (same or paired organ), we considered only the first tumour for the estimation of net survival.

Even with those limits aforementioned, the relative survival setting should be the preferred approach when estimating net survival with population-based data, regardless of the cancer site, because it is less sensitive to inappropriate data changes in comparison to the cause-specific setting. Parametric approaches using flexible regression models for the excess mortality hazard [13,14], could be considered in the case of long follow up time and few cancer deaths, where the Pohar-Perme estimator produces more erratic results.

Authorship contribution

All authors contributed to the manuscript. RS conducted the analysis and the writing under the supervision of LW, BR and AB. BR, LW and AB all reviewed the paper and made final corrections. All authors read and approved the final version of the manuscript.

Conflict of interest statement

None declared.

Funding

This work was supported by the Swiss Cancer League [BIL KFS-3274-08-2013].

Acknowledgements

This collaborative work was supported by the Geneva Cancer Registry and the Cancer Survival Group at London School of Hygiene and Tropical Medicine. The authors would like to thank Elisabetta Rapiti, Christine Bouchardy, Massimo Usel, Gerald Fioretta, Isabelle Neyroud and all collaborators at the Geneva Cancer Registry as well as Michel Coleman and all collaborators at the London School of Hygiene and Tropical Medicine for their support and encouragements.

References

[1] Berkson J, Gage RP. Calculation of survival rates for cancer. 1952.
[2] Ederer F, Cutler SJ. The relative survival rate: a statistical methodology. National cancer institute monograph no. 6. 1961. p. 101–21.
[3] Geskus RB. Data analysis with competing risks and intermediate states. CRC Press; 2015.
[4] Perme MP, Stare J, Estève J. On estimation in relative survival. Biometrics Mar. 2012;68(1):113–20.
[5] Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: the importance of allowing for informative censoring. Stat Med Apr. 2012;31(8):775–86.
[6] Schaffar R, Rachet B, Belot A, Woods L. Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis. Cancer Epidemiol Jun. 2015;39(3):465–72.
[7] Schaffar R, Rapiti E, Rachet B, Woods L. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva Cancer Registry. BMC Cancer Jan. 2013;13(1):609.
[8] Robins J. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. Proc Biopharm Sect Am Stat Assoc 1993:24–33.
[9] Satten G. Estimating the marginal survival function in the presence of time dependent covariates. Stat Probab Lett Oct. 2001; 54(4):397–403.
[10] Cleveland W. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 1979;74:829–36.
[11] Skyrud KD, Bray F, Møller B. A comparison of relative and cause-specific survival by cancer site, age and time since diagnosis. Int J Cancer 2014 Jul 1;135(1):196–203.
[12] Howlader N, Ries LAG, Mariotto AB, Reichman ME, Ruhl J, Cronin KA. Improved estimates of cancer-specific survival rates from population-based data. J Natl Cancer Inst Oct. 2010; 102(20):1584–98.
[13] Remontet L, Bossard N, Belot A, Est J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. no. December 2005. 2007. p. 2214–28.
[14] Charvat H, Remontet L, Bossard N, Roche L, Dejardin O, Rachet B, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. Stat Med Aug. 2016;35(18): 3066–84.

[Chapter Three]

Examining the determinants of long-term excess mortality

In the first chapter, I showed that the reviewed cause of death was preferred for net survival estimation within the cause-specific setting as it uses all available clinical information and was shown to provide more accurate estimates of cause-specific survival.

However, in the second chapter, the subsequent comparison of the two data settings available for net survival estimation shed the light on the superiority of the relative survival setting over the cause-specific setting, even in the presence of reviewed cause of death. The analyses highlighted that the estimation of net survival, which takes into account informative censoring, in the cause-specific setting was sensitive to relatively small changes in the cause of death allocation, whereas the relative survival setting is more robust to non-comparable life tables. The relative survival setting is thus recommended insofar as it is much more robust to violations of the assumptions.

This chapter completes the third aim of the thesis, which is the investigation of the long-term effects of key prognostic factors and treatment for women with breast cancer, using the relative survival setting, which has been shown to be the superior approach.

In order to achieve this aim, I define the following objectives:

- to model long-term excess mortality including clinical variables to evaluate the long-term effect of prognostic factors and
- to assess non-linear and time-varying effects of these factors using the most appropriate methods.

This chapter comprises a paper that was submitted to the *Journal of Clinical Epidemiology* in October 2017, alongside text which describes the background

to the objectives, summarises the approach and findings as well as additional analyses, and which specifies how these publications fulfil the aim and objectives.

Background

Up to this point, univariable methodology has been used to derive accurate estimations of long-term net survival because the purpose has not been the evaluation of the effects of covariables. To do this with univariable methods, stratification of the data would be required. However, as I now wish to consider several covariables concurrently it is unrealistic to sub-divide the data. The alternative strategy is to use multivariable models, which enable the study of the association between cancer patients' excess mortality and an exposure while simultaneously taking into account confounding variables.

Excess hazard models

As previously seen in the Background section, the mortality hazard, for which all deaths are considered as events, can be written as

$$\lambda_o(t) = \lambda_e(t) + \lambda_{excess}(t)$$

where, $\lambda_0(t)$ represents the overall instantaneous mortality hazard rate, which is equal to the sum of the expected instantaneous mortality hazard $\lambda_e(t)$ and the excess instantaneous mortality hazard $\lambda_{excess}(t)$.

$\lambda_{excess}(t)$ is the instantaneous mortality related only to the cancer of interest. When the excess mortality is null, cancer patients experience the same mortality as non-cancer patients who have, in all other ways, the same characteristics. $\lambda_o(t)$ and $\lambda_e(t)$ are considered as known. They are respectively provided by the mortality observed among the cancer patients and by the overall mortality in the population derived from general population life tables. Therefore, in the multivariable framework, it is only the excess mortality hazard rate λ_{excess} that is modelled. We note:

$$\lambda_o(t, X) = \lambda_e(age + t, Z) + \lambda_{excess}(t, X, \beta)$$

where age corresponds to age at diagnosis, t is follow-up time, and $\boldsymbol{Z}$ represents the variables included in the population life table. $\boldsymbol{\beta}$ is the vector of parameters of the model and $\boldsymbol{X}$ express the vector of covariables. It should be noted that $\boldsymbol{X}$ includes the covariables $\boldsymbol{Z}$ as well as age, eventually with additional prognosis factors.

The models which use this additive relationship between mortality rates are considered as "excess hazard models" [96].

Some authors have used a multiplicative relationship between the mortality rates [62,97,98]:

$$\lambda_o(t) = \lambda_{relative}(t) \times \lambda_e(t)$$

Such multiplicative models impose fewer mathematical constraints in comparison to additive models [100]. However, an additive relationship is considered to be more likely in the context of cancer prognosis among a patient population[97,101,102].

Numerous excess hazard models have been developed [97,103]. These have been used across many epidemiological studies and are based on several key assumptions. First, the baseline mortality hazard rate, which is the mortality hazard rate that we would observe when all the variables included in the model are set to their reference value, is assumed to be constant within the specified time intervals. Second, the ratio between the mortality hazard rates of two subgroups of patients is assumed to remain constant over follow-up time. Finally, these models assume that continuous variables have a log-linear effect on the mortality hazard rate.

Flexible excess hazard models

The underlying assumptions of these standard excess hazard models are however restrictive, especially in the context of long-term survival. To relax these, flexible models have recently been proposed for overall survival [104,105] as well as for net survival [106–109]. Their development has been aided by increasing computational power and their purpose is to account for non-linear (NL) and/or time-dependent (TD) effects of covariables using flexible functions such as splines.

Splines

A spline is a mathematical function defined as a combination of different polynomials which are joint at pre-specified point called knots. A spline S takes its value in the interval $[a; b]$ divided in subintervals $[t_{i-1}; t_i]$ with $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$. A polynomial P_i is defined for each interval$[t_{i-1}; t_i]$. We have

$$S(t) = P_1(t), t_0 \leq t < t_1,$$

$$S(t) = P_2(t), t_1 \leq t < t_2,$$

$$\dots$$

$$S(t) = P_k(t), t_{k-1} \leq t < t_k,$$

$$\text{and where } P_j(t_j) = P_{j+1}(t_{j+1}), j = 2, \dots, k-2$$

The given $k+1$ points t_i are defined as knots. The degree of the polynomials as well as the number of knots depend on the complexity of the phenomenon to be modelled. An increasing polynomial's degree or number of knots allows increased flexibility but can also lead to over-adjustment of the data. The position of the knots can be user-defined a priori (using background clinical

knowledge of the phenomenon), or based on the distribution of events in the observed data.

Complex effects

Non-linear effects

A continuous covariable has a linear effect on the mortality hazard rate when the ratio between the mortality rate measured for a covariable X and the mortality rate measured for $X+1$ is independent of X. For example, consider the effect of age on the excess mortality rate. If age at diagnosis has a linear effect, the effect of each increasing year of age is constant across the whole range of ages in the database. That is, the difference in the excess mortality of a 51-year-old compared to a 50-year-old is the same as a 61 compared to a 60-year-old. This hypothesis of linearity is commonly applied but rarely true in practice.

For a specific covariable X, flexible models can relax this hypothesis and are described as follows:

$$\lambda_{excess}(t, X) = f(t) + g(X)$$

The non-linear effect of age is considered with the function $g(X)$. g is a flexible function often described with splines and enables the effect of the variable to be non-linear.

Time-dependent effects

The effect of a covariable on the mortality hazard rate is time dependent (non-proportional) when the ratio between the mortality rate measured for a and the mortality rate measured for $a+1$ is independent from time since diagnosis. In other words, if we are interested in the risk ratio of two subgroups, the risk ratio is

assumed the same at both one and five years after diagnosis; it is considered to be constant through time. Let's imagine an excess hazard model fitting a hazard ratio of 1.84 (Figure 10).

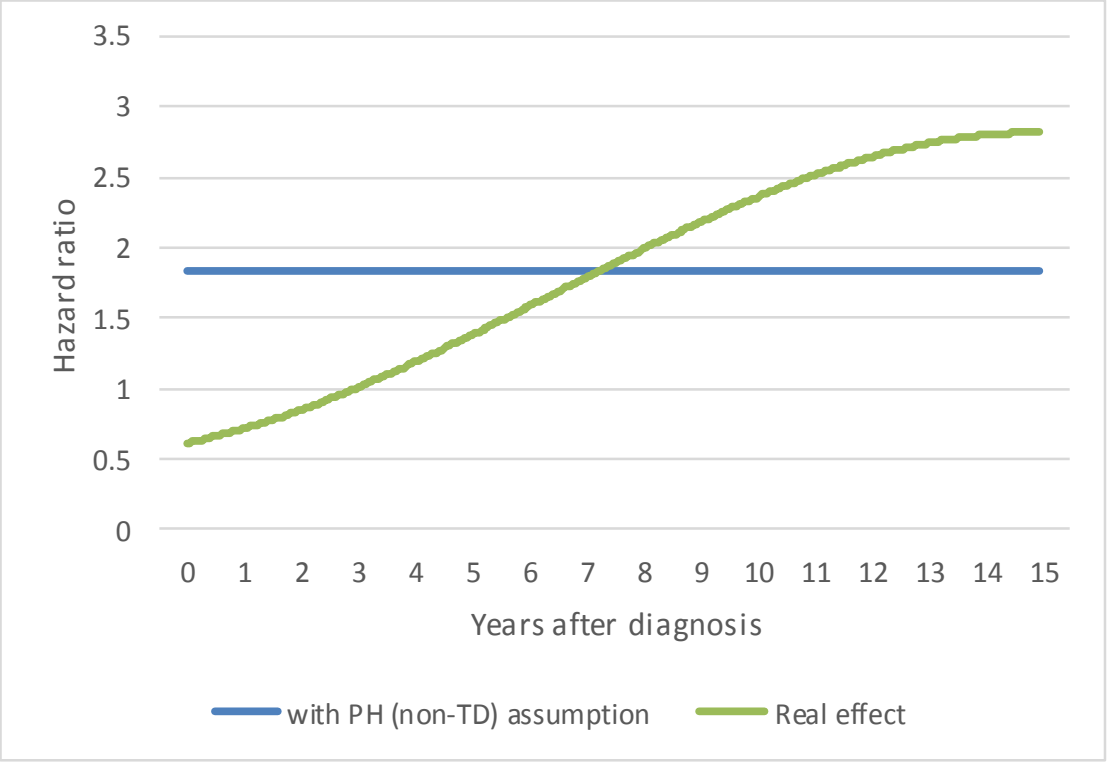


Figure 10: Example of a time-dependent excess hazard ratio in contrast to the assumption of proportional (non-time dependent) effect

Because of the non-time dependent (proportional) assumption the hazard ratio remains constant over time. However, the green line shows that the real hazard ratio varies over time. Similarly, to that of linearity, flexible models can relax this assumption and take into account this time-dependence.

In a time-dependent model we have, for a specific covariable *X*:

$$\lambda_{excess}(t, X) = f(t) + h(t) \times X$$

The time-dependent effect is taken into account by the interaction term $h(t) \times X$ where h represents a spline function of time and enables the effect of X to vary with time since diagnosis.

A single variable X can have both a non-linear and a time dependent effect on the mortality hazard rate. We therefore define:

$$\lambda_{excess}(t, X) = f(t) + g(X) + h(t) \times X$$

Model selection

When performing multivariate analyses, a key consideration is which variables should be included in the model. In the realm of complex effects, we are also interested in testing the presence or absence of significant non-linear or time-dependent effects for each variable. The more covariables we consider, the more models have to be compared.

As described above, several options are available regarding the association between a covariable and excess hazard. We distinguish:

- No effect

  or

  linear (L) or non-linear (NL) effect.

  and/or

- Proportional (P) or time dependent (TD) effect

Each effect needs to be tested independently in order to assess its fit to the observed data. It has been demonstrated that if the shape of an effect is wrongly evaluated it can lead to biased statistical model estimations [110]. In that sense, for a specific covariable, neglecting a non-linear effect could mistakenly

lead to a time-dependent effect being found to be significant for the same variable (and vice versa) [104]. Because of this, specific methods need to be used to decide which variables and which effects are included in the model.

Wynant and Abrahamowicz have proposed a model building strategy for this scenario. It has been proved to be efficient and successful in detecting the correct complex effects as well as eliminating spurious ones [111]. This iterative backward elimination procedure involves testing simultaneously, for each variable, non-linear and time-dependent effects using a "decision tree". This strategy starts from the fitted full model, and successively eliminates spurious non-linear and time-dependent effects.

For example, in the case of a continuous variable X and a categorical variable Y (Figure 11). The full complex model (*Model* C_a) includes all possible complex effects of both covariables (non-linear and time-dependent). The first arm tests whether, for covariable X, the model supports the inclusion of the NL effect, assuming TD effect for all covariables (Figure 11, test 1). The second arm tests whether we could eliminate the TD effects of both X and Y in turn, assuming NL for covariables X (Figure 11, tests 2 and 3). Each of these tests leads to a p-value generated thought the likelihood ratio test (comparison of models 1, 2 and 3 with model C_a). The effect leading to the highest p-value among the 3 tests, if that p-value is higher than 0.05, is discarded. A new full model (Model C_b) is then considered in place of model C_a and tested in turn. This is repeated for models C_c, C_d, …, *etc*. until all effects remain statistically significant. At this point, the model is considered as final (Model F).

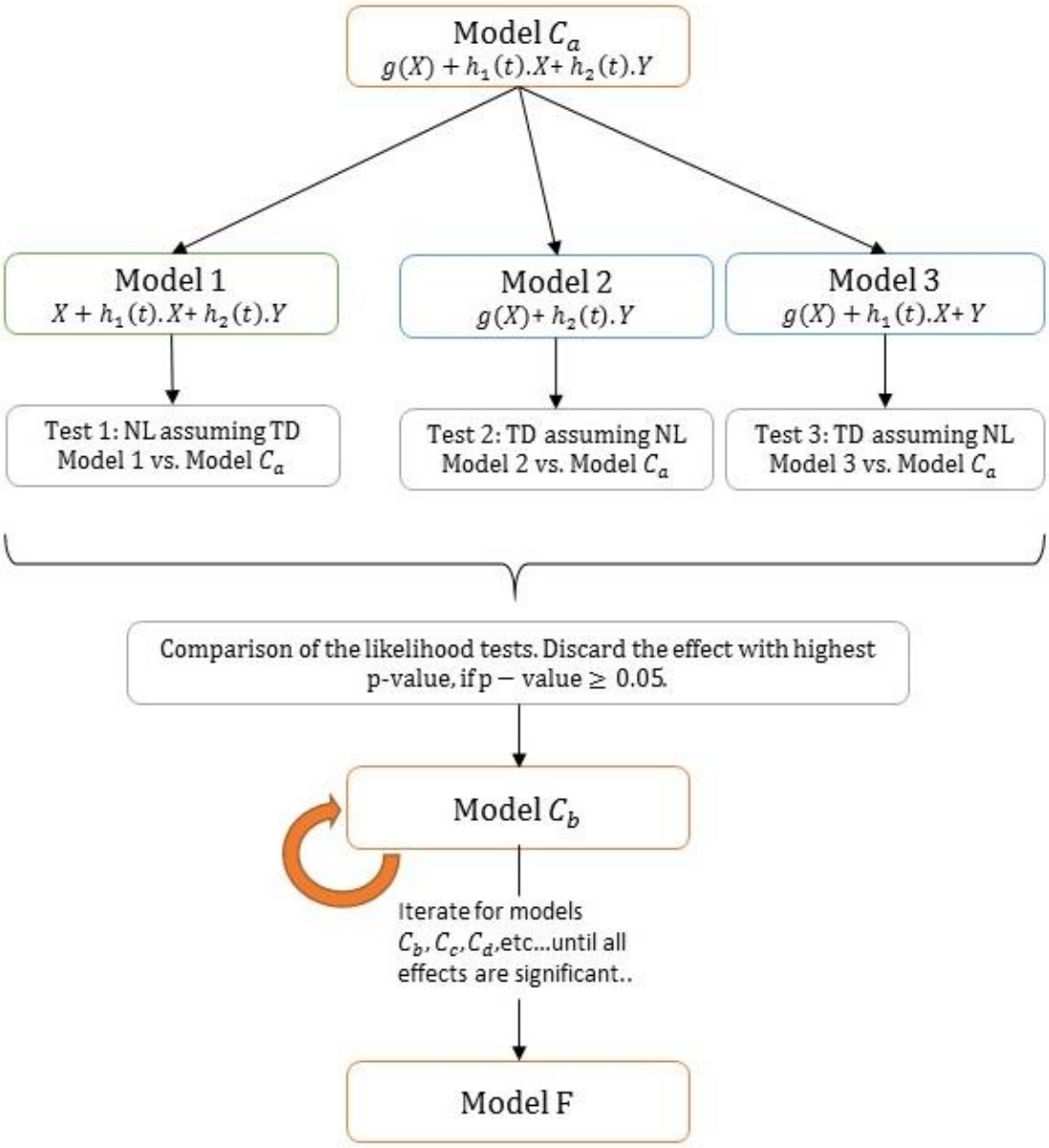


Figure 11: Modelling strategy for selection of covariable X effects. Proposed by Abrahamowicz et al (Wynant and Abrahamowicz, 2014).

Bootstrap analysis

The issue of model stability has been discussed for a very long time in the literature [112,113]. Nevertheless, many studies do not perform a formal evaluation of how well the selected model represents the observed data.

One method of evaluating the stability of the modelling strategy is a bootstrap sensitivity analysis [114]. The analyses are repeated on samples of the same size, drawn with replacement from the initial data, a large number *(B)* of times. When the repetition of the modelling strategy leads systematically to the same results, the selected model can be considered to be robust.

Once these steps (drawing and model selection procedure) have been repeated *B* times the bootstrap inclusion (relative) frequency or BIF can be calculated. This is a measure of the number of times a specific variable/effect is selected in the model selection process over the total of bootstrap sample *B*. Sauerbrei and Schumacher [115] studied the effect of the number of bootstrap replications by varying *B* between 50 and 1000. They concluded that even *B* = 100 can give reasonable results, although in practice they suggested working with several hundred draws.

Paper 3

Modelling long-term excess mortality

Description

Paper 3 presents research which aims to evaluate the association between prognostic factors, treatment and long-term excess mortality for women diagnosed with breast cancer. My modelling strategy considered complex effects: non-linearity and time-dependence, in order to relax underlying assumptions about the pattern of excess mortality, which may be clinically unfounded. I tested the robustness of my results by using bootstrap analyses. All of this was performed using information derived from the Geneva Cancer Registry where detailed population-based data are available. Prognostic factors related to the patients, to the tumours themselves as well as information about treatments given to the women were used in the analyses.

Main results

The final model that I derived suffered from a lack of robustness insofar as covariables and complex effects were very sensitive to the bootstrap analysis. For 60 out of 300 bootstrap samples, the model did not reach convergence. The variables size of tumour, hormone receptors status, age at diagnosis, grade and nodal involvement were more often selected in the sensitivity analysis. However, not all of them were selected in the derived model. Conversely some co-variables/effects that were selected in the derived models were found as being not significant in the sensitivity analyses.

We observed a time-dependent association for *age:* excess mortality increased with age during the first 10 years of follow-up but reversed after this point. Excess mortality increased linearly with *tumour size* and that this association was

constant over time since diagnosis. *Nodal involvement* was associated with higher excess mortality. There was evidence of a TD association for *hormone receptor status*, with negative receptors being associated with an increased risk of dying from breast cancer only during the first 5 years of follow-up. This was similar for *grade*: women with well differentiated tumours displayed a lower risk of dying from breast cancer, an association which also tended towards the null at the end of follow-up. *Radiotherapy* was associated with a decreasing risk of dying during the first 10 years after diagnosis.

Additionally, unexpected results were observed for some covariables. Patients treated with chemotherapy and/or hormonal strategies were associated with an increased risk of dying from breast cancer, likely to be due to indication bias, despite an adjustment for patients and tumours characteristics.

Our results highlighted the importance of taking into account complex effects such as non-linearity and/or time dependence. This was illustrated by the substantial differences between simple model, which assumed linear and non-time dependent effects, and the flexible model I developed. I found some persistent effects for some specific covariables, which have important clinical implications. In particular, time-dependent effects of age and hormone receptors status were observed.

Conclusion

I concluded that an accurate evaluation of the determinants of excess mortality was not possible without further considerations. I observed a gap between the modelling strategy that was theoretically well designed and up-to-date and its application on population-based data. Further analyses should be based on a larger population and use more detailed data (on patients and

tumours characteristics) with long-term follow-up to tackle model instabilities. Furthermore, additional statistical tools from the causal inference framework should be used to address the inherent biases related to observational studies.

Fulfilment of Aims and Objectives

The aim of this chapter was the evaluation of the long-term effect of prognostic factors and treatment on excess mortality due to breast cancer. Although I applied an appropriate methodology to detailed data, I encountered several issues during the analyses, which limited the generalizability of the results outside of this cohort. Despite these limitations, some interesting patterns were observed. In particular, I showed that it was important to consider complex effects, such as time-dependent and non-linear effects when examining long-term excess mortality due to breast cancer.

Lack of robustness

The first issue I encountered is related to the lack of reproducibility of the derived model, which may therefore not be accurate. There are two main reasons this could have arisen.

Insufficient statistical power

Breast cancer has a relatively good prognosis. The number of deaths, which is the event of interest in my study, is therefore not as large amongst breast cancer patients as among other malignancies. In addition, the Geneva population is relatively small, therefore a relatively small number of deaths included in my cohort. Finally, evaluating the complex effects of prognostics factors and treatment on long-term net survival meant that a relatively high number of parameters were considered. Together, this resulted in a low statistical power. Potential strategies to overcome this are considered in the Discussion of Paper 3 and in the Discussion and Perspectives chapter below.

Alternative methodology

It is possible that inappropriate methodology could also have been the underlying reason for the model instabilities. To test whether this was the case, I considered an alternative statistical tool available for the estimation of flexible excess hazard models. This is defined as "*stpm2*" and is an alternative model to "*mexhaz*", which was applied in Paper 3.

"stpm2"

The command *stpm2* is available in Stata [115]. The model was proposed by Royston and Parmar [104] was extended by Lambert and Royston [116]. It is based on an extension of the Weibull distribution through splines and the regression model is defined on the log cumulative hazard scale. According to the authors the advantage of modelling on this scale is that the cumulative hazard, being a function of log time, is stable and easy to transform from survival to hazard and *vice versa*. Complex effects are modelled using restricted cubic splines [117]. The latter use three-degree polynomials, which are forced to be linear beyond the two external knots.

Comparison with "*mexhaz*"

All analyses were performed using the same covariables as those used in my research paper (paper 3), namely age at diagnosis, size of the tumour (mm), nodal involvement (No vs. Yes as reference category), grade of the tumour (Well vs. Moderately/Not differentiated as reference category), hormone receptor status (oestrogen and progesterone, negative vs. positive as reference category), radiotherapy (Yes vs. No as reference category), chemotherapy (Yes vs. No as reference category) and hormonal treatment (Yes vs. No as reference category).

I exactly replicated the analysis based on *mexhaz* reported in Paper 3 using stpm2. Both methods derived almost identical coefficients for all covariables in the simple model, which did not include complex effects (Table 2).

	stpm2		***mexhaz***	
	Coefficient	SD	Coefficient	SD
Age	0.115	0.104	0.114	0.102
Size of the tumour	0.605	0.199	0.664	0.196
Nodes involvment	-0.333	0.191	-0.370	0.185
Grade of the tumour	-0.830	0.311	-0.727	0.288
Hormone receptors	0.455	0.303	0.437	0.297
Radiothereapy	-0.601	0.229	-0.585	0.225
Chemotherapy	0.522	0.247	0.472	0.235
Hormonal treatment	0.285	0.273	0.319	0.267

Table 2: Comparison of the covariables coefficients of the simple model according to the technique used.

However, attempts to fit the most complex *model C,* which included all complex effects, were not comparable. The most complex model derived using *stmp2* did not reach convergence. This prevented any comparison between these two techniques. Had *stpm2* reached convergence, further issues should have needed to be considered. Indeed, it is known that regression models defined on the log cumulative hazard scale are problematic in the context of multiple time-dependent effects. In particular, these models are difficult to present graphically [112,118]. Because my interest was to display long-term association between prognosis factor and the excess mortality hazard in order to bring new insight for long-term survival, it is preferable to use regression models defined on the log-hazard scale, as is done in the *mexhaz* R-package and as presented in Paper 3.

Model misspecification

Besides the issue of model instability, I also observed model misspecification. My study was designed to approach, as far as possible, a randomised control trial by selecting a homogeneous cohort and accounting for several prognostic factors. Additionally, complex effects were considered using a complete modelling strategy in order to relax assumptions that are clinically unlikely. Despite all this, I observed unexpected associations between systemic treatments and excess mortality due to breast cancer.

These observations were almost certainly a result of confounding by indication, whereby difference in excess mortality by treatment may originate from differences in the underlying reasons that these treatments were given, or not, such as presence of comorbidities.

To evaluate this, I performed a stratified analysis restricting the cohort to patients with very similar individual and tumours characteristics. I evaluated long-term net survival by chemotherapy and hormonal treatment. No clear results in favour of the use of chemotherapy nor hormonal treatment were observed (Figure 12, Figure 13). This confirmed that more detailed data are needed to disentangle the effects of treatment. Instrumental variables could have been used but, are not easy to implement in the context of Geneva. There is only one public and university hospital, others are private clinics. Using this as an instrumental variable would have raised additional issues related to socio-economic disparities.

Alternative strategies for dealing with this bias are considered in detail in the Discussion of Paper 3 and in the Discussion and Perspectives chapter below.

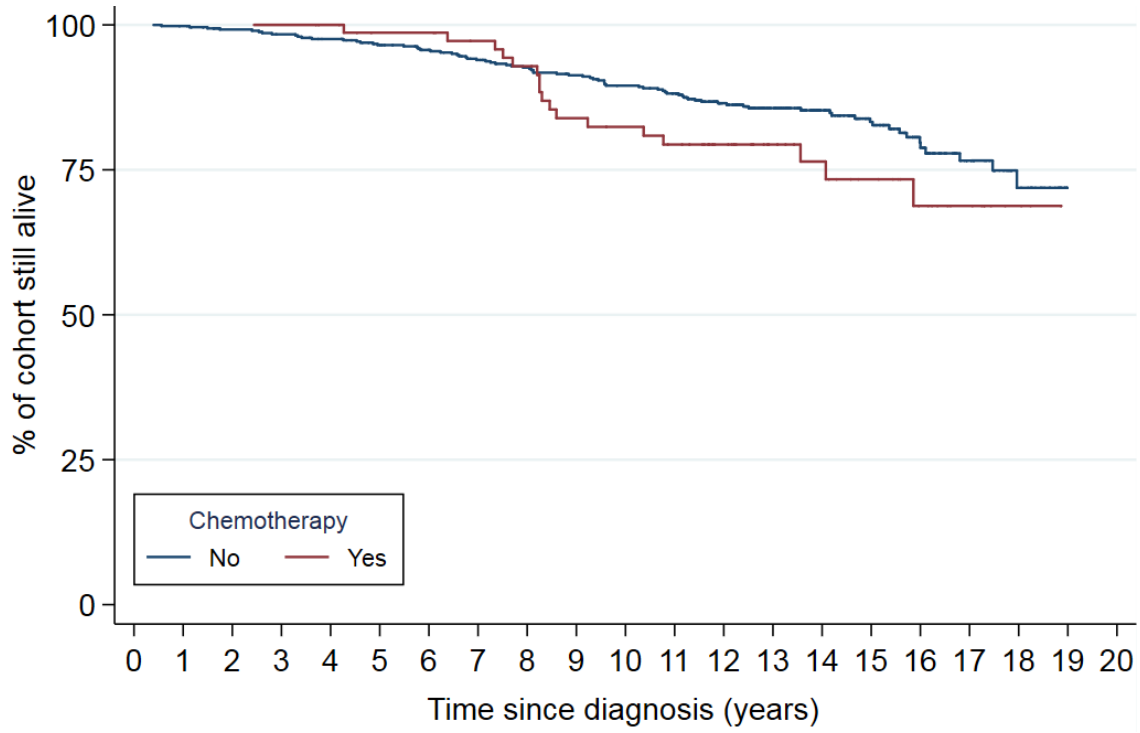


Figure 12: Net survival by chemotherapy for patients diagnosed with breast cancer in Geneva between 1995 and 2002. Patients aged 50-69 with well-differentiated tumours measuring <30mm and with nodal involvement, who also had a least one positive hormonal receptor.

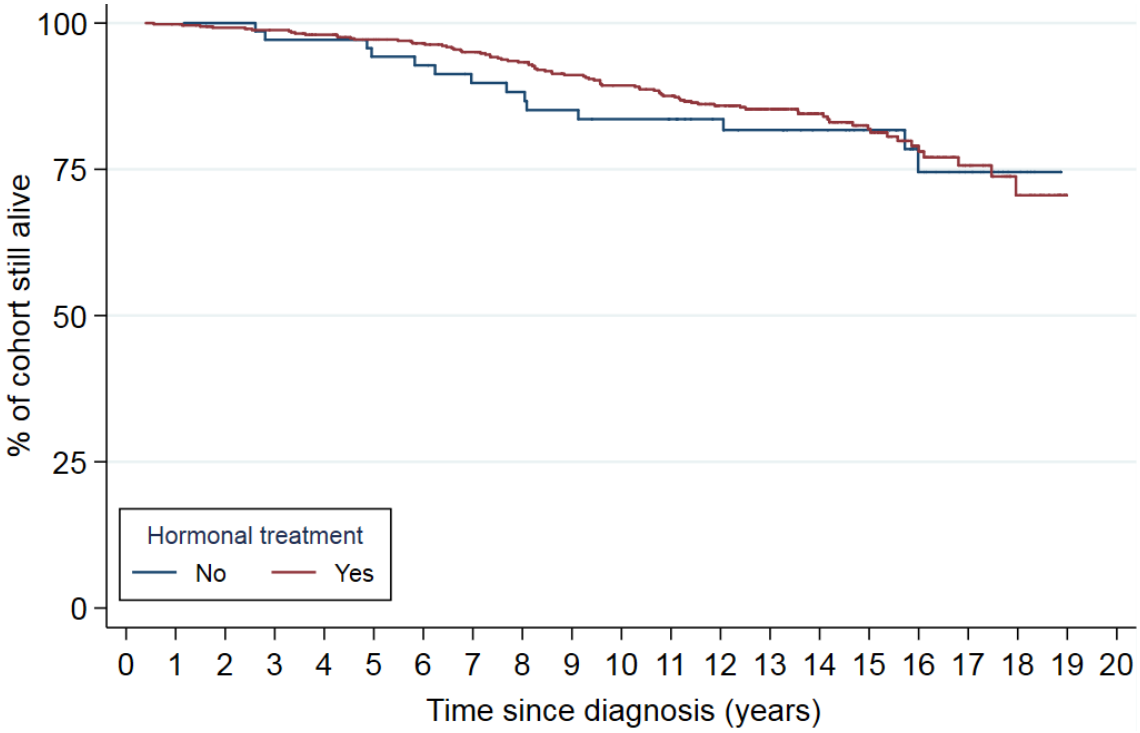


Figure 13: Net survival by hormonal treatment for patients diagnosed with breast cancer in Geneva between 1995 and 2002. Patients aged 50-69 with well differentiated tumours measuring <30mm and with nodal involvement, who also had a least one positive hormonal receptor

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Robin Schaffar
Principal Supervisor	Dr Laura Woods
Thesis Title	Long-term net survival among women diagnosed with breast cancer: accuracy of its estimation and evaluation of its determinants

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*		Was the work subject to academic peer review?	

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Journal of Clinical Epidemiology
Please list the paper's authors in the intended authorship order:	R.Schaffar, A.Belot, B.Rachet, L.Woods
Stage of publication	Rejected in December 2017. New draft in preparation.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was the lead author of this paper. I planned and carried out the literature review and data analysis. I prepared all drafts of the paper. The co-authors provided input and feedback on the content data analysis and on the paper drafts prepared by me.

Student Signature: [redacted] **Date:** 09/03/2018

Supervisor Signature: [redacted] **Date:** 09/03/2018

CAN WE ESTIMATE THE LONG-TERM EFFECT OF PROGNOSTIC FACTORS AND TREATMENT FOR BREAST CANCER USING FLEXIBLE EXCESS HAZARD MODELS IN THE CONTEXT OF CANCER REGISTRY DATA? WOMEN DIAGNOSED 1995-2002 IN GENEVA, SWITZERLAND.

Robin Schaffar[1,2], Aurélien Belot[2], Bernard Rachet[2], Laura Woods[2]

1. Geneva Cancer Registry, University of Geneva, CMU, 1211, Geneva 4, Switzerland

2. Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

INTRODUCTION

Breast cancer is a major disease worldwide. Its prognosis has, however, improved rapidly during the last four decades [1]–[3]. Accordingly, there are increasing numbers of women who have survived breast cancer. Despite this, there is evidence for a lack of population 'cure', that is, the probability of dying as a consequence of the disease persists for many years after diagnosis [4], [5] even for women who are screen-detected [6].

The estimation of net survival has allowed these trends to be observed [7]–[9]. Unlike all other metrics, net survival evaluates the mortality arising only from the disease of interest, disregarding the influence of other causes of death [10]. In the context of long-term survival this is fundamental because the likelihood of death from other causes increases with follow-up time. The use of net survival allows accurate comparisons of patient's subgroups across space and time, between which mortality from other causes may vary considerably [9], [11].

Although there is a great interest, both clinically and epidemiologically, in the determinants of long-term survival for breast cancer patients, follow-up beyond 5 or 10 years has not been widely considered. Indeed, one of the key assumptions of many of the survival models that have been used is the proportionality of the hazard, i.e. that the association between a single covariable and the probability of death is constant through follow-up time. However, it is more probable that the influence of a covariable one year after diagnosis might be different later on.

Several studies have demonstrated such time-varying associations of covariables for breast cancer but very few have been able to consider very long-term follow-up [12], [13]. The small proportion of these studies which have included long-term observations have demonstrated that the associations of some covariables do vary with time since diagnosis in the long term [14]. In particular, the influence of treatment represents an interesting line of investigation since it is likely that certain treatments lead to severe long-term side effects [12], [15].

The Geneva Cancer Registry offers an ideal context to study the evaluation of determinants of long-term net survival. The cancer registry, initiated in 1970, allows very long-term follow-up of cancer patients. The availability of detailed information for each woman's tumour enables multivariable survival analysis.

In this research, we aim to evaluate long-term effects between prognostic factors and the excess mortality hazard for breast cancer patients diagnosed in Geneva, focusing especially on treatment variables. To reach this aim, we focus on early-stage tumours which were surgically resected. We use flexible excess hazard regression models in order to take into account potential time-varying and non-linear associations. We apply a systematic model selection process and check the stability of our model by conducting a sensitivity analysis using bootstrap sampling.

MATERIAL AND METHODS

PATIENT COHORT

The Geneva Cancer Registry collects information on incident cancer cases from various sources, including hospitals, laboratories and private clinics, all of whom report newly diagnosed cancer cases. Trained registrars systematically extract information from the medical records and conduct further investigations in the case of missing data. The registry regularly estimates cancer patient survival, taking as the reference the date the diagnosis was confirmed or, if it preceded the diagnosis and was related to the disease, the date of hospitalisation. In addition to standard examination of death certificates and hospital records, follow-up of the patient's vital status is assessed annually by matching the Registry's database with information held by the Cantonal Population Office who maintain a live register of the resident population.

We included all women diagnosed with an invasive primary breast cancer in the Geneva Canton between 1995 and 2002. We restricted the sample to patients diagnosed with pathological TNM stage I and II disease who were treated with surgery (N=2,029). Among those patients, we excluded patients older than 75 years (N= 232). Information on stage was missing among only 60 (2.57%) patients with surgery. All women were followed-up to 31st December 2013 (minimum of 11 years of follow-up).

PROGNOSTIC FACTORS AND TREATMENT

We focused on established prognostic factors and on treatment. *Age at diagnosis* (years) was included *a priori* as an irrefutable prognostic factor [16]. We considered *tumour size* (mm), *degree of differentiation* (Well vs. Moderately/Not differentiated), *nodal involvement* (No vs. Yes) and *hormone receptor status* (Negative vs. Positive) which together reflect the severity of the disease. We included *radiotherapy*, *chemotherapy* and *hormonal treatment* following surgery (each Yes vs. No) in order to examine the long-term associations of these systemic treatments with survival.

STATISTICAL MODELLING OF THE EXCESS MORTALITY HAZARD

We estimated the excess mortality hazard due to cancer for the patient group. The excess hazard corresponds to the mortality hazard related only to the disease of interest (in our case, breast cancer) and is defined as the difference between the mortality observed amongst a cohort of patients and their expected (background) mortality [Andersen and Vaeth 1989, Esteve 1990]. The association between covariables and excess mortality can vary with time since diagnosis, particularly when considering long-term follow up. For example, a particular treatment might have a strong influence on excess mortality 1 year after diagnosis but a weaker influence 10 years after diagnosis (time-dependent, TD, association). Furthermore, continuous variables can display non-linear (NL) associations (for example, excess mortality might increase exponentially with age). In order to consider these complex associations, we used the flexible excess hazard model proposed by Charvat *et al.* [17], which follows the work of Remontet *et al* [18]. This excess hazard model is implemented in the "*mexhaz*" package written for R software [17], [19].

MODEL BUILDING STRATEGY

We used the model building strategy suggested by Wynant and Abrahamowicz [20]. This iterative backward elimination procedure involves testing, for each variable, the presence of significant TD and, for continuous variables only, NL associations as well as the overall significance of the variable itself. An initial model including all variables, as well as all possible TD and NL associations, is fitted. Potentially spurious NL and TD associations are then eliminated one by one. Our initial model thus included:

- *age at diagnosis* (continuous, NL and TD associations included),
- *tumour size* (continuous, log-transformed, NL and TD associations included),
- *nodal involvement* (binary, TD association included, "Yes" as reference category),
- *grade of the tumour* (binary, TD association included, "Moderately/Not differentiated" as reference category),

- *hormone receptor status* (binary, TD association included, "Positive" as reference category),
- *radiotherapy* (binary, TD association included, "No" as reference category),
- *chemotherapy* (binary, TD association included, "No" as reference category) and
- *hormonal treatment* (binary, TD association included, "No" as reference category).

The model building strategy resulted in a single derived model which included only those variables found to be significant, along with any significant TD and/or NL associations for these variables.

SENSITIVITY ANALYSES

We conducted a sensitivity analysis to examine the stability of the derived model using a bootstrap technique [21]. This involved re-applying the model selection procedure to 300 random samples, drawn, with replacement, from the cohort of cancer patients. This procedure allows the evaluation of the strength of association between a particular covariable and the excess mortality hazard by the calculation of the bootstrap inclusion frequency (BIF). The BIF is the proportion of times a specific variable was included by the model selection process over the total number of samples. We further considered only models where the association was plausible (outliers where estimated values of the Excess Hazard Ratio (HER) were greater than 100 or less than 0.01 were excluded). We then plotted all the estimated functional forms of each covariable ($N \leq 300$), along with the averaged functional form calculated on all the retained samples.

RESULTS

PATIENT COHORT

The study included 1,797 women diagnosed with first primary invasive breast cancer between 1995 and 2002 which was classified as stage I or II at diagnosis and treated surgically (Table 1). Data were missing for at least one co-variable for 12.4% of women. The highest proportion of missing data was for the size of the tumour (N=72, 4.0%). Only women with complete data for all variables were considered for the modelling analyses (N=1,574, 87.6%) [22]. Of these, 351 died (22.3%) and 236 were censored (14.9%) before the end of follow-up. The median follow-up time was 12.8 years.

STABILITY OF THE DERIVED MODEL

The sensitivity analysis, performed to evaluate the performance of the model building strategy, suggested that the model derived for the patient cohort was not very robust. For 60 out of 300 bootstrap samples, the model did not reach convergence. The variables *size of tumour*, *hormone receptors status*, *age at diagnosis*, *grade* and *nodal involvement* displayed the highest BIFs in the sensitivity analysis (Table 3, more than 90%). However, not all of them were selected in the derived model; here neither *nodal involvement* nor *hormone receptor status* showed evidence of an association with the excess mortality hazard. The covariables describing treatment were less frequently selected in the sensitivity analysis, with a BIFs of 75.4%, 59.6% and 45.0% for *chemotherapy*, *radiotherapy* and *hormonal treatment* respectively, whilst in the derived model, *chemotherapy* and a TD association for *hormonal treatment* were retained. Although TD associations were frequently observed in the sensitivity analysis for the covariables *hormone receptors status* and *age* (BIF 95.4% and 87.1% respectively), only the TD association for *age* was found to be significant in the derived model. NL associations for *age* and *size of the tumour* were not retained in the derived model, which was consistent with the low BIFs observed in the sensitivity analysis (18.8% and 34.2% respectively).

VARIABLE ASSOCIATIONS WITHIN THE SENSITIVITY ANALYSIS

Figures 1, 2 and 3 display the associations between each of the covariables and excess mortality derived from the sensitivity analysis, without outliers. The mean association across all samples (black solid line) is also displayed. These show that within the sensitivity analysis we observed a TD association for *age:* excess mortality increased with age during the first 10 years of follow-up (Figure 1, a-b) but reversed after this point (Figure 1, c). Figure 2 shows that excess mortality increased linearly with *tumour size* and that this association was constant over time since diagnosis. *Nodal involvement* was associated with higher excess mortality. There was evidence of a TD association for *hormone receptor status*, with negative receptors being associated with an increased risk of dying from breast cancer only during the first 5 years of follow-up. This was similar for *grade*: women with well differentiated tumours displayed a lower risk of dying from breast cancer, an association which also tended towards the null at the end of follow-up. *Radiotherapy* was associated with a decreasing risk of dying during the first 10 years after diagnosis, whereas receipt of *chemotherapy* and *hormonal treatment* were associated with an increasing risk during the entire follow-up period.

DISCUSSION

The determinants of long-term survival are currently of particular interest because of the dramatic increase in the number of patients surviving breast cancer matched to the observation that these women are never 'cured'. Understanding the impact of prognostic factors and of treatment with time since diagnosis is therefore increasingly important. In this context, population-based data are crucial to understand the influence of treatment and outcomes for all cancer patients.

Our approach

In order to estimate the long-term effects of prognostic factors and treatment on the risk of dying from breast cancer, we used data from the population-based Geneva Cancer Registry. Randomised clinical trials are the gold standard for evaluating the effect of a treatment. However, RCTs only include highly selected groups of patients who do not represent the general population of cancer patients. Moreover, the assessment of the benefit is mostly done at a relatively short term [23]. In order to estimate these long-term effects using observational data, we restricted our cohort to a relatively homogeneous group of younger patients (less than 75) with localised disease (stage I and II) and who had received surgery. We adjusted for several covariables indicating the severity of the disease, as well as taking in account differences in individual characteristics. Furthermore, we considered flexible excess hazard models to estimate the mortality related to the disease after controlling for other causes. We considered non-linear and time-dependent associations to relax the assumptions that the association of continuous variables is linear and that the EHRs of all variables are constant throughout time since diagnosis. Both of these assumptions are clinically unlikely in the context of long-term survival. We used a recommended strategy [20] for selection of covariables and their complex associations, and performed a sensitivity analysis to evaluate the reproducibility of the model [21].

Despite using, on a fairly homogeneous group of patients, this optimised and up-to-date modelling strategy, a clear process for variable and complex association selection and a sensitivity analysis, our results demonstrated a lack of stability and model misspecification, associated with unrealistic effects of some treatments.

Modelling issues

First, our sensitivity analysis demonstrated that the set of covariables included (eventually with NL and/or TD functional forms) to model the excess mortality hazard was unstable. Because of this demonstrated instability, results obtained from a single model should be interpreted with caution. This is best illustrated by the fact that a significant proportion of models (20%) did not reach convergence during the sensitivity analysis, as well as the fact that several variables selected for the single derived model were rarely retained in the sensitivity analysis (low BIF). Meanwhile others not retained in the derived model were often selected by the sensitivity analysis (high BIF).

There are a number of possible reasons for this lack of robustness. The first is related to the context in which the study was conducted. Since breast cancer patients present with high survival, the number of events is relatively low in breast cancer data, even given long-term follow-up. This is especially true for Geneva, which has a fairly small population (495,000 inhabitants), and in the studied population restricted to early stage cancer patient. It is recommended that at least 5 or 10 events per parameter should be included when estimating regression coefficients [24], [25]. Because we considered both time-dependent and non-linear associations for all prognostic variables, the number of parameters included in our model was large relative to the number of deaths. The convergence issues that we encountered are therefore likely to be explained, in part, by a lack of power. However, decreasing the number of parameters (either by reducing the number of variables, or excluding some complex associations) would not have been a better strategy, given that our core aim was to try to better understand the long-term associations of prognostic covariables for breast cancer patients.

Neither was it practical to increase the number of women in order to increase the number of events since this could only have been done by including women with advanced disease, for which treatment protocols are very different, or by including elderly women, who do not have the opportunity for long-term follow-up.

The analysis excluded 12.3% of the cohort because of missing data, thus leading to a loss of information. However this proportion is relatively low for these types of observational data and complete-case analyses have been proved to be sufficiently efficient for such ranges of missing data proportion [22].

It is possible the lack of stability may have been a result of the modelling approach. We consider this unlikely, however. The flexible regression model we applied has been purposefully designed to estimate excess mortality hazard and take into account complex associations. The model selection strategy has previously been shown to be efficient and successful in detecting the correct complex associations as well as eliminating spurious ones [20].

The second main issue was that our strategy was unable to fully control for confounding by indication leading to model misspecification. This would be an issue even with a perfectly robust model. This confounding is best illustrated by the unexpected results for chemotherapy and hormonal treatment. Women receiving these treatments experienced an increased risk of dying from breast cancer compared to women who did not receive them (Figure 3). This reflects the fact that the patients in the cohort who received chemotherapy and hormonal treatment were those with more advanced disease at diagnosis (Table 2). This represents a limitation of our strategy, which was not able to account for the fact that almost all women who were likely to benefit from these therapies were given them, resulting in a small or absent comparison group within the patient cohort (confounding by indication). We performed a stratified analysis to explore this (data not shown). We grouped patients with very similar characteristics together and compared their survival according to receipt of chemotherapy or not. This similarly showed an increased risk in the excess hazard of death associated with chemotherapy. This strongly

suggests that additional information about the prognosis of patients not receiving chemotherapy is missing from our dataset, and that this led to misspecification of the model.

In addition, interactions between treatment received and other co-variables might be required. Although we planned to examine the existence of such interactions, they were tricky to implement due to the convergence issues we encountered during the modelling process, and not reasonable to explore in our small sample size dataset.

Other possible strategies

Our results point towards the need for different statistical strategies in addition to our modelling strategy to be better able to examine these associations. Causal inference analyses would be one suitable approach [26]–[28]. The objective of causal inference is to mimic the randomised trial that would have been set for the research question by using observational data and specific statistical techniques. This would require much more detailed data on comorbidities and other factors used to define the treatment choice. Furthermore, software to implement causal inference techniques is not yet available for the excess mortality hazard. Further methodological research is thus required to enable such analyses to be conducted.

Clinical interpretations

Nevertheless, a few cautious clinical interpretations can be drawn from these data. Some co-variables presented high BIFs within the sensitivity analysis and the observed associations appeared stable to the exclusion of outliers suggesting that they are indicative of a robust, underlying associations. Consistent with Jatoi *et al.* [13] we found that patients with negative hormone receptors presented a higher excess mortality during the first years after diagnosis compared to those who have positive hormone receptors (BIF 95.4%). Regarding age at diagnosis, our results corroborated exactly with those found by Cluze *et al.* [16] which showed the risk of dying from breast cancer was associated with increasing age at 1 and 5 years after diagnosis but that this association reversed at 10 years (BIF 87.1%). In addition to hormone

receptor status and age at diagnosis, tumour size, grade and nodal involvement displayed associations which were similar to those described in a previous meta-analysis [14]. Although our results are broadly consistent with previous studies, caution should be exercised in reporting the size of these associations, given that they have been derived from models which display a lack of robustness. We observed a time-dependent association for radiotherapy: patients treated with radiotherapy exhibited a decreased risk of dying in the first 10 years following their diagnosis but an increased risk afterwards. This association was, however, sensitive to the inclusion or exclusion of outliers. That said, it could potentially correspond to late side effects of treatment, in particular cardiac complications, which are known as a likely consequence of irradiations given close to the heart [29]–[31].

Conclusion

Our research aimed to estimate the long-term effects of prognostic factors and treatment for breast cancer using flexible excess hazard models for patients diagnosed in Geneva between 1995 and 2002. Our study highlights the challenges of interpreting these associations in observational data and as well as the need for high quality and detailed clinical information at a population level so that these associations can be examined in detail. With such data, causal inference methods could be applied to be able to describe an effect rather than an association. However, applying causal inference methods requires further methodological work and the development of specialist software for the use of causal inference in the context of excess hazard modelling.

ACKNOWLEDGMENTS

This collaborative work was conducted with the support of the Geneva Cancer Registry and the Cancer Survival Group at London School of Hygiene and Tropical Medicine. We should like to thank Elisabetta Rapiti, Isabelle Neyroud, Massimo Usel, Christine Bouchardy and all collaborators at the Geneva Cancer Registry as well as Michel Coleman and all collaborators at the London School of Hygiene and Tropical Medicine for their support and encouragement.

BIBLIOGRAPHY

[1] A. Gondos, F. Bray, T. Hakulinen, H. Brenner, and EUNICE Survival Working Group, "Trends in cancer survival in 11 European populations from 1990 to 2009: a model-based analysis," *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 20, no. 3, pp. 564–573, Mar. 2009.

[2] A. Gondos *et al.*, "Recent trends in cancer survival across Europe between 2000 and 2004: a model-based period analysis from 12 cancer registries," *Eur. J. Cancer Oxf. Engl. 1990*, vol. 44, no. 10, pp. 1463–1475, Jul. 2008.

[3] M. Sant, S. Francisci, R. Capocaccia, A. Verdecchia, C. Allemani, and F. Berrino, "Time trends of breast cancer survival in Europe in relation to incidence and mortality," *Int. J. Cancer*, vol. 119, no. 10, pp. 2417–2422, Nov. 2006.

[4] L. M. Woods, B. Rachet, P. C. Lambert, and M. P. Coleman, "'Cure' from breast cancer among two populations of women followed for 23 years after diagnosis," *Ann. Oncol.*, vol. 20, no. 8, pp. 1331–1336, Aug. 2009.

[5] H. Brenner and T. Hakulinen, "Are patients diagnosed with breast cancer before age 50 years ever cured?," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 22, no. 3, pp. 432–8, Feb. 2004.

[6] L. M. Woods, M. Morris, and B. Rachet, "No 'cure' within 12 years of diagnosis among breast cancer patients who are diagnosed via mammographic screening: women diagnosed in the West Midlands region of England 1989–2011," *Ann. Oncol.*, vol. 27, no. 11, pp. 2025–2031, Nov. 2016.

[7] N. Bossard *et al.*, "Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM).," *Eur. J. Cancer Oxf. Engl. 1990*, vol. 43, no. 1, pp. 149–60, Jan. 2007.

[8] E. Crocetti, V. De Lisi, L. Gafà, O. Sechi, and L. Mangone, "Net survival: comparison between relative and cause-specific survival estimates.," *Epidemiol. Prev.*, vol. 25, no. 3 Suppl, pp. 32–6, Jan. 2001.

[9] Z. Uhry, N. Bossard, L. Remontet, J. Iwaz, L. Roche, and GRELL EUROCARE-5 Working Group and the CENSUR Working Survival Group, "New insights into survival trend analyses in cancer population-based studies: the SUDCAN methodology," *Eur. J. Cancer Prev. Off. J. Eur. Cancer Prev. Organ. ECP*, vol. 26 Trends in cancer net survival in six European Latin Countries: the SUDCAN study, pp. S9–S15, Jan. 2017.

[10] M. Pohar Perme, J. Estève, and B. Rachet, "Analysing population-based cancer survival - settling the controversies," *BMC Cancer*, vol. 16, no. 1, p. 933, Dec. 2016.

[11] C. Allemani *et al.*, "Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2)," *Lancet Lond. Engl.*, vol. 385, no. 9972, pp. 977–1010, Mar. 2015.

[12] I. Jatoi *et al.*, "Time-Varying Effects of Breast Cancer Adjuvant Systemic Therapy," *J. Natl. Cancer Inst.*, vol. 108, no. 1, p. djv304, 2015.

[13] I. Jatoi, W. F. Anderson, J.-H. Jeong, and C. K. Redmond, "Breast cancer adjuvant therapy: time to consider its time-dependent effects.," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 29, no. 17, pp. 2301–4, Jun. 2011.

[14] I. Soerjomataram, M. W. J. Louwman, J. G. Ribot, J. a Roukema, and J. W. W. Coebergh, "An overview of prognostic factors for long-term survivors of breast cancer.," *Breast Cancer Res. Treat.*, vol. 107, no. 3, pp. 309–30, Feb. 2008.

[15] B. Bodai, "Breast Cancer Survivorship: A Comprehensive Review of Long-Term Medical Issues and Lifestyle Recommendations," *Perm. J.*, vol. 19, no. 2, Apr. 2015.

[16] C. Cluze *et al.*, "Analysis of the effect of age on the prognosis of breast cancer.," *Breast Cancer Res. Treat.*, vol. 117, no. 1, pp. 121–9, Sep. 2009.

[17] H. Charvat *et al.*, "A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates," *Stat. Med.*, vol. 35, no. 18, pp. 3066–3084, Aug. 2016.

[18] L. Remontet, N. Bossard, A. Belot, J. Estève, and French network of cancer registries FRANCIM, "An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies," *Stat. Med.*, vol. 26, no. 10, pp. 2214–2228, May 2007.
[19] R Development Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2009.
[20] W. Wynant and M. Abrahamowicz, "Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis.," *Stat. Med.*, vol. 33, no. 19, pp. 3318–37, Aug. 2014.
[21] P. Royston and W. Sauerbrei, *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester, England ; Hoboken, NJ: John Wiley, 2008.
[22] A. Marshall, D. G. Altman, P. Royston, and R. L. Holder, "Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study," *BMC Med. Res. Methodol.*, vol. 10, p. 7, Jan. 2010.
[23] J. Warwick, L. Tabàr, B. Vitak, and S. W. Duffy, "Time-dependent effects on survival in breast carcinoma: results of 20 years of follow-up from the Swedish Two-County Study," *Cancer*, vol. 100, no. 7, pp. 1331–1336, Apr. 2004.
[24] J. Concato, P. Peduzzi, T. R. Holford, and A. R. Feinstein, "Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy," *J. Clin. Epidemiol.*, vol. 48, no. 12, pp. 1495–1501, Dec. 1995.
[25] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford, "Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates," *J. Clin. Epidemiol.*, vol. 48, no. 12, pp. 1503–1510, Dec. 1995.
[26] M. A. Hernán and J. M. Robins, "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available," *Am. J. Epidemiol.*, vol. 183, no. 8, pp. 758–764, Apr. 2016.
[27] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.," *Stat. Sci.*, vol. 5, no. 4, pp. 465–472, 1990.
[28] S. Wright, *Correlation and Causation*. 1921.
[29] Y.-J. Cheng *et al.*, "Long-Term Cardiovascular Risk After Radiotherapy in Women With Breast Cancer," *J. Am. Heart Assoc.*, vol. 6, no. 5, May 2017.
[30] L. M. Boerman *et al.*, "Long-term outcome of cardiac function in a population-based cohort of breast cancer survivors: A cross-sectional study," *Eur. J. Cancer Oxf. Engl. 1990*, vol. 81, pp. 56–65, Jun. 2017.
[31] K. Rygiel, "Cardiotoxic effects of radiotherapy and strategies to reduce them in patients with breast cancer: An overview," *J. Cancer Res. Ther.*, vol. 13, no. 2, pp. 186–192, Jun. 2017.

Figure 1: Excess hazard ratio for age at diagnosis, excluding outliers, using 70 years as the reference (a) 1 year after diagnosis. (b) 5 years after diagnosis. (c) 10 years after diagnosis.

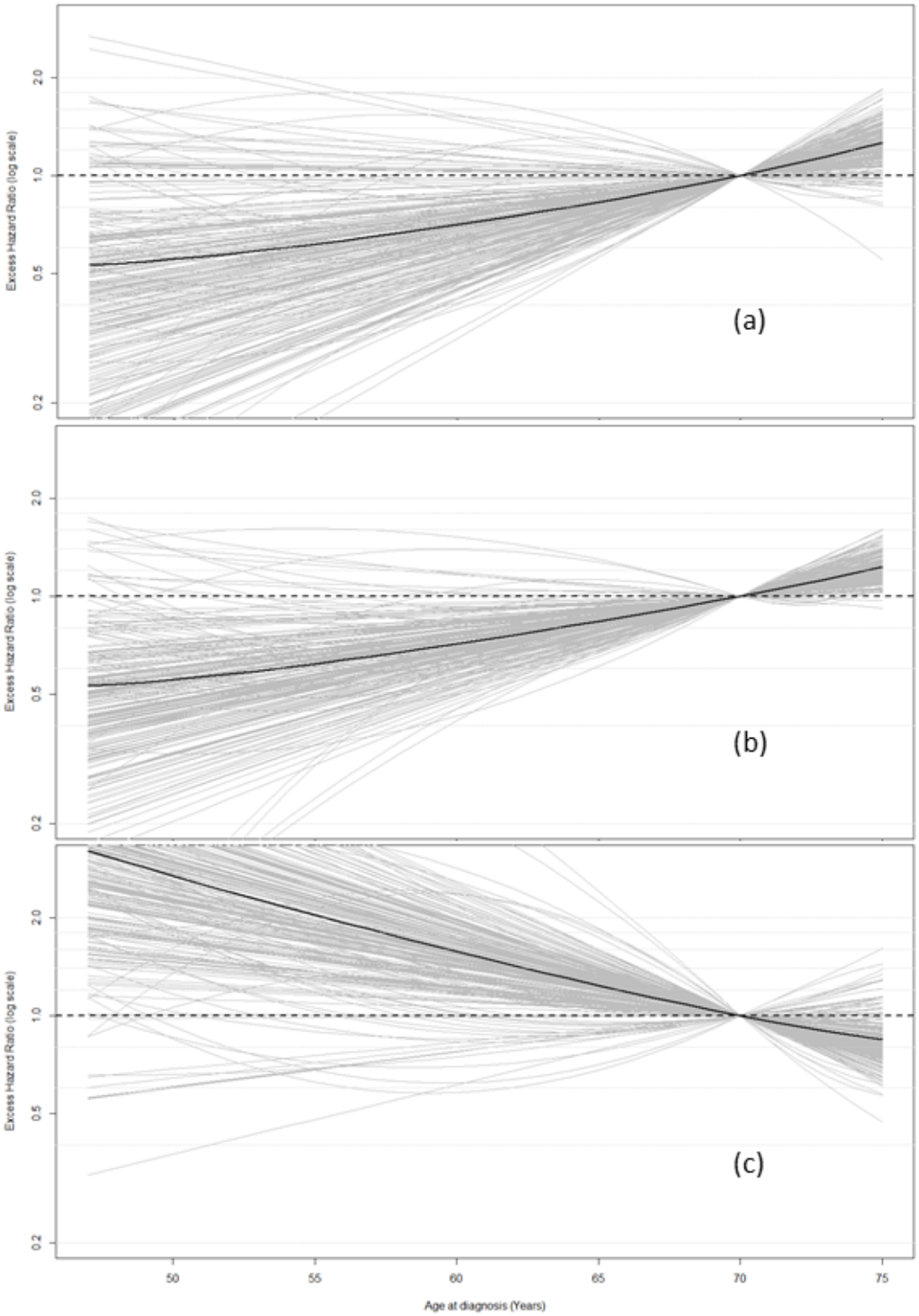


Figure 1: Excess hazard ratio for tumour size , excluding outliers, using 20 mm as a reference. (a) 1 year after diagnosis (b) 5 year after diagnosis (c) 10 year after diagnosis.

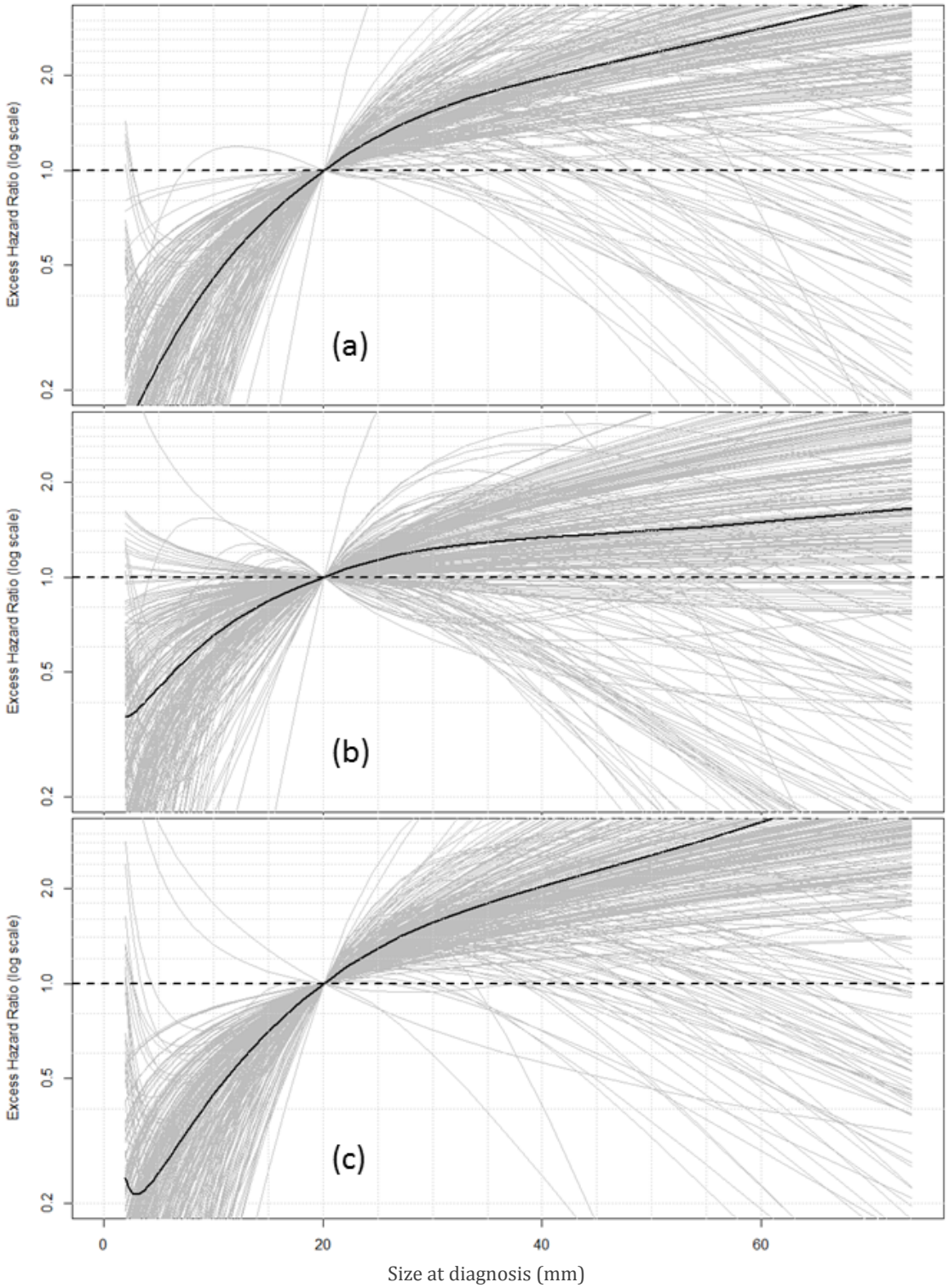


Figure 1: Excess hazard ratio for categorical covariables, excluding outliers. (a) Nodal involvement with Yes as reference category. (b) Grade with Moderately/Not differentiated as reference category. (c) Hormone receptor status with Positive as reference category. (d) Radiotherapy with No as reference category. (e) Chemotherapy with No as reference category. (f) Hormonal treatment with No as reference category.

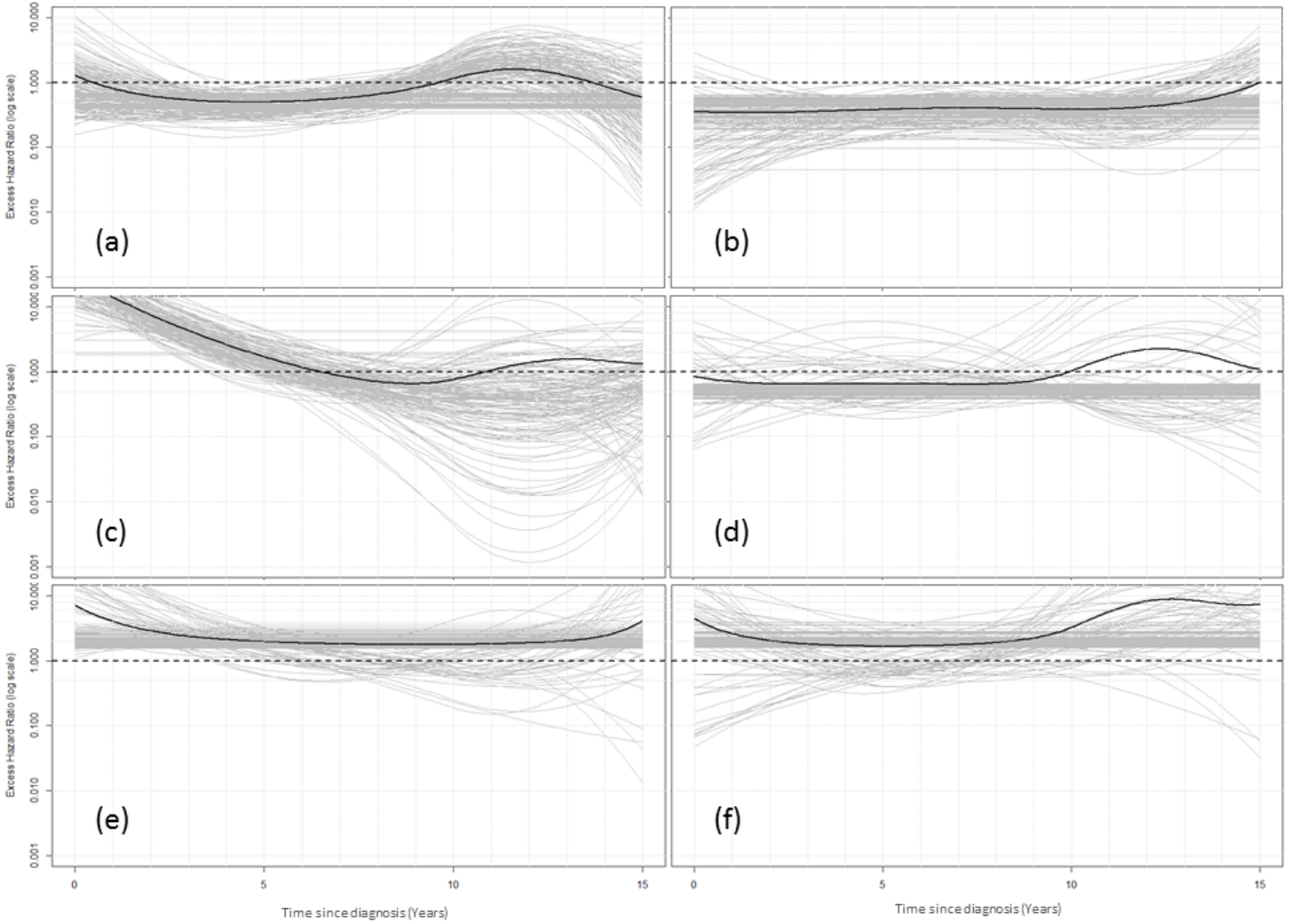


Table 1: Characteristics of the patients diagnosed with breast cancer between 1995 and 2002.

	N	%
Age group		
<40	87	4.8
40-49	359	20.0
50-59	652	36.3
60-69	535	29.8
70-79	164	9.1
Total	1,797	100.0
Size in mm		
0-9	271	15.1
10-19	800	44.5
20-29	395	22.0
30-39	160	8.9
40+	99	5.5
Missing	72	4.0
Total	1,797	100.0
Nodal involvement		
N+	500	27.8
N0	1,265	70.4
Total	1,797	100.0
Differentiation		
Well differentiated	1,113	61.9
Moderately/ poorly differentiated	617	34.3
Missing	67	3.7
Total	1,797	100.0
Hormone receptors		
Positive	1,520	84.6
Negative	212	11.8
Missing	65	3.6
Total	1,797	100.0
Radiotherapy		
No	291	16.2
Yes	1,506	83.8
Total	1,797	100.0
Chemotherapy		
No	1,006	56.0
Yes	791	44.0
Total	1,797	100.0
Hormonal treatment		
No	492	27.4
Yes	1,305	72.6
Total	1,797	100.0

Table 1: Patient and tumour characteristics according to treatment.

	Radiotherapy				Chemotherapy				Hormonal treatment			
	No		Yes		No		Yes		No		Yes	
	N	%	N	%	N	%	N	%	N	%	N	%
Age group												
<40	16	5.5	71	4.7	14	1.4	73	9.2	39	7.9	48	3.7
40-49	79	27.1	280	18.6	125	12.4	234	29.6	149	30.3	210	16.1
50-59	78	26.8	574	38.1	348	34.6	304	38.4	158	32.1	494	37.9
60-69	87	29.9	448	29.7	374	37.2	161	20.4	114	23.2	421	32.3
70-79	31	10.7	133	8.8	145	14.4	19	2.4	32	6.5	132	10.1
Total	291	100	1506	100	1006	100	791	100	492	100	1305	100
Size in mm												
0-9	44	15.1	227	15.1	222	22.1	49	6.2	84	17.1	187	14.3
10-19	91	31.3	709	47.1	512	50.9	288	36.4	172	35.0	628	48.1
20-29	66	22.7	329	21.8	170	16.9	225	28.4	120	24.4	275	21.1
30-39	45	15.5	115	7.6	56	5.6	104	13.1	51	10.4	109	8.4
40+	27	9.3	72	4.8	33	3.3	66	8.3	38	7.7	61	4.7
Missing	18	6.2	54	3.6	13	1.3	59	7.5	27	5.5	45	3.4
Total	291	100	1506	100	1006	100	791	100	492	100	1305	100
Nodal involvement												
N+	82	28.2	418	27.8	127	12.6	373	47.2	137	27.8	363	27.8
N0	201	69.1	1064	70.7	863	85.8	402	50.8	351	71.3	914	70.0
Missing	8	2.7	24	1.6	16	1.6	16	2.0	4	0.8	28	2.1
Total	291	100	1506	100	1006	100	791	100	492	100	1305	100
Differentiation												
Well differentiated	186	63.9	927	61.6	497	49.4	616	77.9	346	70.3	767	58.8
Moderately/ poorly differentiated	84	28.9	533	35.4	472	46.9	145	18.3	114	23.2	503	38.5
Missing	21	7.2	46	3.1	37	3.7	30	3.8	32	6.5	35	2.7
Total	291	100	1506	100	1006	100	791	100	492	100	1305	100
Hormone receptors												
Positive	220	75.6	1300	86.3	920	91.5	600	75.9	253	51.4	1267	97.1
Negative	38	13.1	174	11.6	45	4.5	167	21.1	194	39.4	18	1.4
Missing	33	11.3	32	2.1	41	4.1	24	3.0	45	9.1	20	1.5
Total	291	100	1506	100	1006	100	791	100	492	100	1305	100
Complete data												
Complete	216	74.2	1358	90.2	900	89.5	674	85.2	397	80.7	1177	90.2
Missing	75	25.8	148	9.8	106	10.5	117	14.8	95	19.3	128	9.8
Total	291	100	1506	100	1006	100	791	100	492	100	1305	100

Table 1: Bootstrap Inclusion Frequency (BIF) for each co-variable and their type of associations following the sensitivity analysis.

	BIF (%)		
	Main	Non-linear	Time dependent
Age	92.9	45.0	87.1
Size of the tumour	99.6	82.0	57.1
Nodes involvment	85.8	-	55.4
Grade of the tumour	90.8	-	51.7
Hormone receptors	97.1	-	95.4
Radiotherapy	59.6	-	18.8
Chemotherapy	75.4	-	32.9
Hormonal treatment	45.0	-	20.4

"-": Not applicable.

[Discussion and perspectives]

Summary

An increasing number of women are surviving breast cancer and so a better understanding of the long-term consequences of the disease is needed. This thesis evaluated first, the most accurate way to estimate long-term net survival when reliable cause of death information is available, and second, what can be determined about the long-term effects of prognostic factors and treatment for women with breast cancer. Two main messages can be drawn from this research. First, the type of data setting used to derive net survival is crucial to obtain a robust estimation. Second, data currently available in cancer registries requires recently developed methodologies to allow a relevant clinical evaluation of the determinants of the long-term net survival.

First research question

What is the most accurate way to estimate net survival when reliable cause of death information is available?

These analyses have shown that the relative survival setting was the less biased setting for net survival estimation, particularly for long-term survival, even in the presence of validated cause of death.

An accurate comparison

In the context of population based data, the perception that "relative survival" was superior to "cause-specific survival" has been pervasive for a number of years in the dedicated literature [68,119–121]. It thus may appear that these conclusions are not novel. Sarfati *et al.* for instance compared relative survival and cause-specific survival and demonstrated that relative survival was the less biased [122]. Like us, they studied the biases related to the data: misclassification

for the cause-specific survival and non-comparability of life tables for relative survival. Misclassification of cause of death has also frequently been considered. Amongst others, Percy *et al.* has demonstrated several times that misclassification was an issue in the context of the estimation of mortality [70,123,124] The work from Percy *et al.* [70] has been cited almost 800 times in the literature.

However, all these studies considered "cause-specific survival" and "relative survival" to both be measuring survival related to the disease of interest; none accounted for informative censoring. This bias occurs when cancer patients are removed from the risk set in a non-random way. My research has thus added a further aspect to these methodological developments. I have conducted, for the first time, an accurate comparison between the two data settings available for the estimation of net survival taking into account the bias of informative censoring for both settings. This research is therefore unique and shows that, when comparing two theoretically unbiased estimators of net survival, the relative survival setting is, indeed, less sensitive to violations of the assumptions (non-comparability between the cohort of cancer patients and the group used to derive their expected mortality compared to misclassification of the cause of death for the cause-specific setting).

The results described in this thesis are therefore progress those described in previous work. In addition to the work of Pohar-Perme *et al.*, this work provides a comparison between the estimator proposed in the relative survival setting and an unbiased estimator in the cause-specific setting and has demonstrated the superiority of the relative survival setting for the estimation of net survival. These results concur with those of Percy *et al.*, in demonstrating that misclassification

of the underlying cause of death is also an issue when estimating net survival as well as mortality. Finally, in addition to the work from Sarfati *et al.*, this work demonstrates that the relative survival setting was less biased in the context of net survival. This research highlights the importance of being aware of both the bias of informative censoring and the difference between the two data settings when interpreting survival estimates in the population-based setting.

Evaluation of long-term net survival

A further contribution made by this thesis is the extension of the results to long-term net survival. It has been demonstrated that misclassification of the cause of death tends to increase with time since diagnosis. This is probably because determining that a death is due to breast cancer is likely to be easier a few months after diagnosis than decades later, as well as related to the fact that the numbers of deaths decrease with time. This observation further strengthens the recommendation of the relative survival setting for the estimation of net survival, even when validated cause of death is known, but especially when it is not.

Application across cancer sites

The superiority of the relative-survival setting was demonstrated for breast cancer but also for three other cancer sites. Although I did not examine all cancer sites, I selected a full range of different types of cancer. This allowed the examination of whether the results were robust to different contexts. On the basis of these analyses, it is hard to imagine a very specific cancer site for which the cause-specific setting would prevail over the relative-survival one. This could however be tested in further research studies by simply running similar analyses on all cancer sites.

Reviewed cause of death

Because the Geneva Cancer Registry performs clerical review of the officially recorded cause of death from death certificates, reviewing it using all the clinical information available, these data offered a unique opportunity to evaluate its validity. This represents one of the key strength of this research and enabled us to show that, even in the presence of high quality data, the relative survival setting is less biased for the estimation of net survival.

Life tables

Like the reviewed cause of death represented a strength in my examination of the cause-specific setting, the complete and smoothed life tables I used are an advantage in the examination of the relative-survival setting. It was the first time recommendations from Rachet *et al.* [90] were applied to data from Geneva. This increased the validity of the mortality rates used for the expected rate of death and thus of the estimation of net survival.

Implications and perspectives

Implications for cancer registries

These results have wide-ranging implications for the standardisation of survival analysis worldwide. This is important for geographical comparisons as well as for temporal and subgroup comparisons at a local level. The use of a common method would ensure that observed results are not due to either biases related to the type of data or in background mortality. There are already some examples of large studies which use the unbiased estimator in the relative survival setting for the estimation of net survival [125–127]. My research should add weight to the importance of using the best methodology for these comparisons, especially if examining long-term survival.

My research should also motivate registries worldwide to unite in using only the relative survival setting for the estimation of net survival for their own areas. There may be a residual reluctance about the advantage of this because the size of bias itself is often moderate [128–130]. However, a better understanding of this estimator as the only consistent method should motivate common consensus. Moreover, it is important to communicate that this method is robust especially for long-term survival, which has not been previously demonstrated.

The work presented in this thesis also implies that the accessibility of the statistical tool itself should be improved. The rapid implementation of this is important since the current statistical software available to derive net survival using the Pohar-Perme estimator is not as user-friendly as other methods have been.

Implications for clinicians

More specifically, it is important to communicate the superiority of the relative survival setting against the use of the cause of death among clinicians. Indeed clinicians may be resistant to adopting a method, which does not rely on cause of death. Perhaps because of their individual interactions with patients, analyses based on cause of death tends to appear inherently more trustworthy and concrete to clinical specialists. This research should therefore be shared beyond cancer registry employees to all clinicians and collaborators using their data for research, in order to raise awareness of relative survival setting being more robust for the estimation of net survival.

Cost-effectiveness of validated cause of death

One important question arising from these results is whether, for the Geneva Cancer Registry, spending resources on reviewing the cause of death is useful.

I have shown that cause of death is not required for the estimation of net survival. It is, however, used for the estimation of mortality rates. For this, reviewed cause of death could presumably be considered as the gold standard. However, in Geneva, official cause of death is used for the estimation of mortality in order to ensure comparability with the rest of Switzerland. The reviewed cause of death is therefore not used for mortality statistics calculated by the Cancer Registry.

From a practical perspective, reviewing a cause of death constitutes a heavy and time-consuming workload. Originally, in the 1970s, Registrars in Geneva noted differences between the cause appended on the death certificate and that in the medical files. They therefore implemented a reviewed cause of death variable in addition to the official one. Now, with the 3,000 incident cases per year, reviewing the cause of death represents the equivalent of a half-time registrar, more than 10% of the total time spent on cancer registration. The cost-effectiveness of reviewed cause of death could thus be evaluated.

Second research question

When using an accurate approach, what can be determined about the long-term effects of prognostic factors and treatment for women with breast cancer?

Despite using a well-designed modelling strategy and very detailed data, I encountered two key difficulties in developing a comprehensive understanding of long-term determinants of excess mortality due to breast cancer in my cohort. These were instability and misspecification of the model. These two difficulties could be overcome by using datasets with a greater number of

women, and so inflating the number of deaths, as well as more detailed data and additional methodologies, in order to adjust for confounding by indication.

Context of the study

As stated in chapter 3, Geneva has a fairly small population but a well-established, long-running Cancer Registry. Each year, approximately 3,000 invasive cancers are diagnosed in the canton, from which 400 are breast cancers. As a result, the number of deaths due to breast cancer is small, but the data do enable long-term follow-up of all women, as well as providing detailed clinical data on their cancers.

Since a better population to conduct this analysis in would need to be both larger but also retain the same long-term follow-up in addition to a high level of detail for the data, the number of Cancer Registries, which could be used for this purpose, is relatively small. Very few other registries have long-term detailed data for larger populations. Collaborative studies would be a logical way to inflate the number of events in order to increase the likelihood of producing a stable model, however detailed long-term data would not necessarily be comprehensively available. For example, Switzerland is covered by 20 cancer registries but half of them were established after 2005 meaning that long-term net survival would be difficult to calculate. Because of the small population covered and the network available between health actors, detailed data are easily available in the Geneva Cancer Registry. This, however, is less likely to be the case in in a larger area across different registry areas or between countries.

Several collaborative studies have already been initiated at European (EUROCARE [126,131–135]) or International level (CONCORD [125,136,137]) but the level of detail remain insufficient for the purpose of my research. As an illustration, from

the covariables we used in my research, only age and grade were requested by the last call from Eurocare-6. TNM stage, size of tumour, or treatment were asked only if available, as not all cancer registries record these data. Information on hormone receptor status was not requested. Additionally, important differences may exist between these databases leading to problems when using them in combination. For example, treatment and screening strategies are different across regions or countries and should be addressed specifically in the analysis and resulting interpretations.

Alternatively, several cancer registries in Europe, such those covering Netherlands or Nordic countries may have the data required for this type of research. Indeed, they are old enough to provide long follow-up of cancer patients and are known to collect clinical data for a sufficient number of patients [138–141]. A comprehensive survey of data available within these registries would enable the feasibility of conducting further research to be established.

It should be added that data that are more precise have now entered clinical practice. Genetic profiles and recent progress in treatment strategies, such us immunotherapies, local radiotherapies, and new hormonal treatments were not considered in my research because no long-term data is available for these yet. These innovations and their accurate recording will however allow an overall improvement in the data quality and potentially help to disentangle the effects of long-term determinants of excess mortality due to breast cancer.

Methodology

In standard epidemiological study designs, statistical power is calculated prior to the commencement of the study. However, multivariable time-to-event analysis with consideration of complex effects is a context in which sample size

analysis is a challenging task [142]. No adequate studies were found dealing with this purpose. Moreover, it is important to note that the convergence issues encountered were revealed during the sensitivity analyses and therefore with sampled data. These particular data, because they were drawn with replacement, could have been significantly different from the baseline data. In contrast, the single model deriving from baseline data did actually converge. Hence, sample size calculations, if they had been possible to implement, would have been unlikely to detect the need for additional events.

I also considered whether the lack of reproducibility arose because of the use of an inappropriate methodology. I considered one alternative model (as described in chapter 3) in order to test whether the issue of non-convergence originated from computational difficulties with the *mexhaz* command. These analyses showed, however, that *mexhaz* provided the most convincing results. I observed that the other statistical tool "*stpm2*" was difficult to fit and an accurate comparison between them was therefore not possible. No additional comparisons were performed but further research could be dedicated to a formal comparison of other tools available for excess hazard modelling. "*mexhaz*" was designed specifically for this purpose and I have shown that it is adequate for this research question.

Further research

Despite the challenges encountered my work was able to highlight how crucial it is to take into account non-linear and non-proportional effects, especially in the context of long-term excess mortality. I also demonstrated persistent effects for some covariables all of which were consistent with what has previously been described in the literature. These results suggest that the underlying patterns of

excess mortality related to these covariables were successfully identified in my models, but that further research is required.

Even without issues of model reproducibility, the evaluation of the effects of prognostic factors and treatment on long-term excess mortality related to breast cancer are likely to have been problematic in the context of these data. A perfectly robust model would have been able to describe associations but in order to evaluate causal effects, additional methodologies must be considered. This is illustrated by the unexpected results that were found for treatment covariables, which are likely to have arisen as a result of confounding by indication. This is because other factors, unmeasured and unaccounted for, play a very influential role in the association within the data.

Additional methods with more detailed data would complement what has been already performed with this research. These additional methodologies could come from causal inference, the objective of which is to mimic a targeted trial and evaluate causal effect between exposure and outcome in the context of observational data.

Propensity scores [143], which belong to the sphere of causal inference, could for example be implemented within the modelling strategy we developed for the research question. Their purpose is to emulate randomization for treatment attribution by creating a score, which is the individual probability of having the treatment in respect to a set of pre-specified covariables.

[Conclusion]

Breast cancer remains a major concern worldwide. The fight against this disease, in order to improve patients' life and monitoring involves a large spectrum of research areas. With this thesis, I have provided new guidance for epidemiologists on how to study long-term net survival and its determinants in the context of population-based data. Taken together my research has presented a strong case for the strength of cancer registries in bringing knowledge regarding long-term survival of breast cancer patients.

Thanks to the high data quality in Geneva we were able to demonstrate that the relative survival setting was the best option for the estimation of long-term net survival. This was true despite the availability of reviewed cause of death. I demonstrated that the findings can be extended to other cancer localisations.

My research has also shown that despite the great interest around the increasing number of breast cancer survivors, it remains difficult to establish with certainty the long-term determinants of deaths related to the disease for these particular patients. We believe that cancer registries are key players for this purpose but further considerations are needed. Long-term follow up data need to be gathered and collected at a higher level of detail for a larger population. Additional methodologies are required to tackle inherent biases described for observational studies. These methodologies need to be developed in the context of excess hazard modelling.

While there are limitations to the use of cancer registry data, as is the case with most data, these are outweighed by the value of building on recent experience to close gaps in cancer knowledge that are difficult, if not impossible, to address with other approaches.

Cochran, 1972

"…observational studies are an interesting and challenging field which demands a good deal of humility, since we can claim only to be groping toward the truth."

[Bibliography]

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. *GLOBOCAN 2012 v1.1, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11*. (2014).
2. Forouzanfar, M. H. *et al.* Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *Lancet* **378,** 1461–1484 (2011).
3. Insistute for health metrics and evaluation, University of Washington. The challenge ahead: Progress and setbacks in breast and cervical cancer. (2011).
4. Adami, H. O., Malker, B., Rutqvist, L. E., Persson, I. & Ries, L. Temporal trends in breast cancer survival in Sweden: significant improvement in 20 years. *J. Natl. Cancer Inst.* **76,** 653–659 (1986).
5. Sant, M. *et al.* Time trends of breast cancer survival in Europe in relation to incidence and mortality. *Int. J. Cancer* **119,** 2417–2422 (2006).
6. Gondos, A., Bray, F., Hakulinen, T., Brenner, H. & EUNICE Survival Working Group. Trends in cancer survival in 11 European populations from 1990 to 2009: a model-based analysis. *Ann. Oncol.* **20,** 564–573 (2009).
7. Gondos, A. *et al.* Recent trends in cancer survival across Europe between 2000 and 2004: a model-based period analysis from 12 cancer registries. *Eur. J. Cancer* **44,** 1463–1475 (2008).
8. Vorobiof, D. A. Recent advances in the medical treatment of breast cancer. *F1000Research* **5,** 2786 (2016).
9. Woods, L. M., Morris, M. & Rachet, B. No 'cure' within 12 years of diagnosis among breast cancer patients who are diagnosed via mammographic screening: women diagnosed in the West Midlands region of England 1989–2011. *Annals of Oncology* **27,** 2025–2031 (2016).

10. Woods, L. M., Rachet, B., Lambert, P. C. & Coleman, M. P. 'Cure' from breast cancer among two populations of women followed for 23 years after diagnosis. *Annals of Oncology* **20,** 1331–1336 (2009).
11. Woods, L. M., Rachet, B., Cooper, N. & Coleman, M. P. Predicted trends in long-term breast cancer survival in England and Wales. *Br. J. Cancer* **96,** 1135–1138 (2007).
12. Brenner, H. & Hakulinen, T. Are patients diagnosed with breast cancer before age 50 years ever cured? *J. Clin. Oncol.* **22,** 432–438 (2004).
13. Singhal, M. K. & Raina, V. Cure from breast cancer, not quite yet but getting there? *Annals of Oncology* **20,** 1291–1292 (2009).
14. Janssen-Heijnen, M. L. G. *et al.* Small but significant excess mortality compared with the general population for long-term survivors of breast cancer in the Netherlands. *Annals of Oncology* **25,** 64–68 (2014).
15. Ambrogi, F., Trevisi, L., Martelli, G. & Boracchi, P. Is breast cancer curable: a study of long-term crude cumulative incidence. *Tumori* **100,** 406–414 (2014).
16. Clèries, R. *et al.* Estimating long-term crude probability of death among young breast cancer patients: a Bayesian approach. *Tumori Journal* **102,** 555–561 (2016).
17. Louwman, W. J., Klokman, W. J. & Coebergh, J. W. Excess mortality from breast cancer 20 years after diagnosis when life expectancy is normal. *Br. J. Cancer* **84,** 700–703 (2001).
18. Brinkley, D. & Haybittle, J. L. Long-term survival of women with breast cancer. *Lancet* **1,** 1118 (1984).
19. Bodai, B. Breast Cancer Survivorship: A Comprehensive Review of Long-Term Medical Issues and Lifestyle Recommendations. *The Permanente Journal* **19,** (2015).

20. Colzani, E. *et al.* Prognosis of patients with breast cancer: causes of death and effects of time since diagnosis, age, and tumor characteristics. *J. Clin. Oncol.* **29,** 4014–4021 (2011).
21. Bellera, C. A. *et al.* Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med Res Methodol* **10,** 20 (2010).
22. Hilsenbeck, S. G. *et al.* Time-dependence of hazard ratios for prognostic factors in primary breast cancer. *Breast Cancer Res. Treat.* **52,** 227–237 (1998).
23. Natarajan, L. *et al.* Time-varying effects of prognostic factors associated with disease-free survival in breast cancer. *Am. J. Epidemiol.* **169,** 1463–1470 (2009).
24. Jatoi, I., Anderson, W. F., Jeong, J.-H. & Redmond, C. K. Breast cancer adjuvant therapy: time to consider its time-dependent effects. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **29,** 2301–4 (2011).
25. Jatoi, I. *et al.* Time-Varying Effects of Breast Cancer Adjuvant Systemic Therapy. *J. Natl. Cancer Inst.* **108,** (2016).
26. Warwick, J., Tabàr, L., Vitak, B. & Duffy, S. W. Time-dependent effects on survival in breast carcinoma: results of 20 years of follow-up from the Swedish Two-County Study. *Cancer* **100,** 1331–1336 (2004).
27. Zahl, P. H. & Tretli, S. Long-term survival of breast cancer in Norway by age and clinical stage. *Statistics in medicine* **16,** 1435–49 (1997).
28. Cooper, G. M. *The cell: a molecular approach*. (ASM Press [u.a.], 2000).
29. Vargas, A., López, M., Lillo, C. & Vargas, M. J. [The Edwin Smith papyrus in the history of medicine]. *Rev Med Chil* **140,** 1357–1362 (2012).

30. Zarshenas, M. M. & Mohammadi-Bardbori, A. A medieval description of metastatic breast cancer; from Avicenna's view point. *Breast* **31,** 20–21 (2017).
31. Drake, R. L., Vogl, W., Mitchell, A. W. M. & Gray, H. *Gray's anatomy for students*. (Churchill Livingstone/Elsevier, 2015).
32. *DeVita, Hellman, and Rosenberg's Cancer Principles & Practice of Oncology: Hardbound Two-Volume Set Plus Integrated Content Website*.
33. Lee, Y. T. Breast carcinoma: pattern of metastasis at autopsy. *J Surg Oncol* **23,** 175–180 (1983).
34. Coleman, C. Early Detection and Screening for Breast Cancer. *Semin Oncol Nurs* (2017). doi:10.1016/j.soncn.2017.02.009
35. McPherson, K. ABC of breast diseases: Breast cancer---epidemiology, risk factors, and genetics. *BMJ* **321,** 624–628 (2000).
36. Edge, S. B. & Compton, C. C. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* **17,** 1471–1474 (2010).
37. Abrams, G. *Neoplasia II*. (2009).
38. Adami, H. O., Malker, B., Holmberg, L., Persson, I. & Stone, B. The relation between survival and age at diagnosis in breast cancer. *N. Engl. J. Med.* **315,** 559–563 (1986).
39. Anders, C. K. *et al.* Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *J. Clin. Oncol.* **26,** 3324–3330 (2008).
40. Eaker, S., Dickman, P. W., Bergkvist, L., Holmberg, L. & Uppsala/Orebro Breast Cancer Group. Differences in management of older women

influence breast cancer survival: results from a population-based database in Sweden. *PLoS Med.* **3,** e25 (2006).

41. Woods, L. M., Rachet, B. & Coleman, M. P. Origins of socio-economic inequalities in cancer survival: a review. *Ann. Oncol.* **17,** 5–19 (2006).
42. Lundqvist, A., Andersson, E., Ahlberg, I., Nilbert, M. & Gerdtham, U. Socioeconomic inequalities in breast cancer incidence and mortality in Europe-a systematic review and meta-analysis. *Eur J Public Health* **26,** 804–813 (2016).
43. Cronin-Fenton, D. P. *et al.* Comorbidity and survival of Danish breast cancer patients from 1995 to 2005. *British Journal of Cancer* **96,** 1462–1468 (2007).
44. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* **40,** 373–383 (1987).
45. Nelson, S. H. *et al.* Impact of very low physical activity, BMI, and comorbidities on mortality among breast cancer survivors. *Breast Cancer Res. Treat.* **155,** 551–557 (2016).
46. Shen, Y. *et al.* Role of detection method in predicting breast cancer survival: analysis of randomized screening trials. *J. Natl. Cancer Inst.* **97,** 1195–1203 (2005).
47. Duffy, S. W. *et al.* Correcting for Lead Time and Length Bias in Estimating the Effect of Screen Detection on Cancer Survival. *American Journal of Epidemiology* **168,** 98–104 (2008).
48. Woods, L. M., Rachet, B., O'Connell, D. L., Lawrence, G. & Coleman, M. P. Are international differences in breast cancer survival between Australia and the UK present amongst both screen-detected women and non-

screen-detected women? survival estimates for women diagnosed in West Midlands and New South Wales 1997-2006. *Int. J. Cancer* **138,** 2404–2414 (2016).

49. Woods, L. M., Rachet, B., O'Connell, D., Lawrence, G. & Coleman, M. P. Impact of deprivation on breast cancer survival among women eligible for mammographic screening in the West Midlands (UK) and New South Wales (Australia): Women diagnosed 1997-2006. *Int. J. Cancer* **138,** 2396–2403 (2016).
50. Sellers, A. H. The clinical classification of malignant tumours: the TNM system. *Canadian Medical Association journal* **105,** 836 passim (1971).
51. Fisher, E. R. *et al.* Fifteen-year prognostic discriminants for invasive breast carcinoma: National Surgical Adjuvant Breast and Bowel Project Protocol-06. *Cancer* **91,** 1679–1687 (2001).
52. Hatteville, L., Mahe, C. & Hill, C. Prediction of the long-term survival in breast cancer patients according to the present oncological status. *Stat Med* **21,** 2345–2354 (2002).
53. Tai, P. *et al.* Short- and long-term cause-specific survival of patients with inflammatory breast cancer. *BMC Cancer* **5,** 137 (2005).
54. Soerjomataram, I., Louwman, M. W. J., Ribot, J. G., Roukema, J. a & Coebergh, J. W. W. An overview of prognostic factors for long-term survivors of breast cancer. *Breast cancer research and treatment* **107,** 309–30 (2008).
55. Mouridsen, H. T., Rose, C., Brodie, A. H. & Smith, I. E. Challenges in the endocrine management of breast cancer. *Breast* **12 Suppl 2,** S2-19 (2003).
56. Giuliano, A. E. *et al.* Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual: Updates to

the AJCC Breast TNM Staging System: The 8th Edition. *CA: A Cancer Journal for Clinicians* **67,** 290–303 (2017).

57. Booth, C. & Tannock, I. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *BJC* 551–555 (2014). doi:10.1038/bjc.2013.725
58. Dos Santos Silva, I. The role of cancer registries. in *Cancer epidemiology: Principles and methods* (International Agency for research on cancer).
59. Parkin, D. M. The role of cancer registries in cancer control. *Int. J. Clin. Oncol.* **13,** 102–111 (2008).
60. Ellis, L. *et al.* Cancer incidence, survival and mortality: explaining the concepts. *Int. J. Cancer* **135,** 1774–1782 (2014).
61. Böhmer, P. E. *Theorie der unabhängigen Wahrscheinlichkeiten Rapports*. 327–343 (1912).
62. Berkson, J. & Gage, R. P. Calculation of survival rates for cancer. *Proc Staff Meet Mayo Clin* **25,** 270–286 (1950).
63. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53,** 457 (1958).
64. Pohar Perme, M., Estève, J. & Rachet, B. Analysing population-based cancer survival - settling the controversies. *BMC Cancer* **16,** 933 (2016).
65. Eloranta, S. *et al.* How can we make cancer survival statistics more useful for patients and clinicians: an illustration using localized prostate cancer in Sweden. *Cancer Causes Control* **24,** 505–515 (2013).
66. Kalbfleisch, J. D. & Prentice, R. L. The Statistical Analysis of Failure Data. *IEEE Transactions on Reliability* **35,** 11–11 (1986).
67. Ederer, F., Axtell, L. M. & Cutler, S. J. The relative survival rate: a statistical methodology. *Natl Cancer Inst Monogr* **6,** 101–121 (1961).

68. Gamel, J. W. & Vogel, R. L. Non-parametric comparison of relative versus cause-specific survival in Surveillance, Epidemiology and End Results (SEER) programme breast cancer patients. *Stat Methods Med Res* **10,** 339–352 (2001).

69. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34,** 187–220 (1972).

70. Percy, C., Stanek, E. & Gloeckler, L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American journal of public health* **71,** 242–50 (1981).

71. Schaffar, R., Rapiti, E., Rachet, B. & Woods, L. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva Cancer Registry. *BMC cancer* **13,** 609 (2013).

72. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. (2011).

73. Gooley, T. A., Leisenring, W., Crowley, J. & Storer, B. E. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med* **18,** 695–706 (1999).

74. Satagopan, J. M. *et al.* A note on competing risks in survival data analysis. *Br. J. Cancer* **91,** 1229–1235 (2004).

75. Prentice, R. L. *et al.* The analysis of failure times in the presence of competing risks. *Biometrics* **34,** 541–554 (1978).

76. Ederer, F. & Cutler, S. J. The relative survival rate: A statistical methodology. in *National cancer institute monograph no.* 6 101–121 (1961).

77. Ederer, F. & Heise,H. Instructions to IBM 650 programmers in processing survival computations. Methodological note no. 10. *End Results Evaluation Section* **National Cancer Institute,** MD (1959).

78. Hakulinen, T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* **38,** 933–942 (1982).

79. Perme, M. P., Stare, J. & Estève, J. On Estimation in Relative Survival. *Biometrics* (2011). doi:10.1111/j.1541-0420.2011.01640.x

80. Robins, J. Information recovery and bias adjustement in proportional hazards regression analysis of randomized trials using surrogate markers. *Proceedings of the Biopharmmaceutical Section, American Statistical Association* (1993).

81. Satten, G. Estimating the marginal survival function in the presence of time dependent covariates. *Statistics & Probability Letters* **54,** 397–403 (2001).

82. Ezzati, M. & Lopez, A. D. Estimates of global mortality attributable to smoking in 2000. *Lancet* **362,** 847–852 (2003).

83. Dickman, P. W. & Adami, H.-O. Interpreting trends in cancer patient survival. *J. Intern. Med.* **260,** 103–117 (2006).

84. Ellis, L., Coleman, M. P. & Rachet, B. The impact of life tables adjusted for smoking on the socioeconomic difference in net survival for laryngeal and lung cancer. *Under review* (2014).

85. Warburton, D. E. R. & Bredin, S. S. D. Health benefits of physical activity: a systematic review of current systematic reviews. *Curr. Opin. Cardiol.* **32,** 541–556 (2017).

86. Kushi, L. & Giovannucci, E. Dietary fat and cancer. *Am. J. Med.* **113 Suppl 9B,** 63S–70S (2002).

87. Key, T. J., Allen, N. E., Spencer, E. A. & Travis, R. C. The effect of diet on risk of cancer. *Lancet* **360,** 861–868 (2002).

88. Hutter, R. V. Cancer prevention and detection. Status report and future prospects. *Cancer* **61,** 2372–2378 (1988).

89. Blakely, T., Sarfati, D. & Shaw, C. What proportion of cancer is due to obesity? *N. Z. Med. J.* **122,** 9–13 (2009).

90. Rachet, B. *et al.* Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health* **15,** 1240 (2015).

91. Ewbank, D. C., Gomez De Leon, J. C. & Stoto, M. A. A reducible four-parameter system of model life tables. *Popul Stud (Camb)* **37,** 105–127 (1983).

92. Elandt-Johnson, R. C. & Johnson, N. L. *Survival Models and Data Analysis*. (John Wiley & Sons, 1980).

93. Kostaki, A. A relational technique for estimating the age-specific mortality pattern from grouped data. *Mathematical Population Studies* **9,** 83–95 (2000).

94. Schaffar, R., Rachet, B., Belot, A. & Woods, L. Cause-specific or relative survival setting to estimate population-based net survival from cancer? An empirical evaluation using women diagnosed with breast cancer in Geneva between 1981 and 1991 and followed for 20 years after diagnosis. *Cancer epidemiology* **39,** 465–72 (2015).

95. Schaffar, R., Rachet, B., Belot, A. & Woods, L. M. Estimation of net survival for cancer patients: Relative survival setting more robust to some assumption violations than cause-specific setting, a sensitivity analysis on empirical data. *European Journal of Cancer* **72,** 78–83 (2017).

96. Estève, J., Benhamou, E., Croasdale, M. & Raymond, L. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med* **9,** 529–538 (1990).

97. Hill, C., Laplanche, A. & Rezvani, A. Comparison of the mortality of a cohort with the mortality of a reference population in a prognostic study. *Stat Med* **4,** 295–302 (1985).

98. Andersen, P. K. & Vaeth, M. Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* **45,** 523–535 (1989).

99. Stare, J., Henderson, R. & Pohar, M. An individual measure of relative survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54,** 115–126 (2005).

100. Buckley, J. D. Additive and multiplicative models for relative survival rates. *Biometrics* **40,** 51–62 (1984).

101. Bolard, P., Quantin, C., Esteve, J., Faivre, J. & Abrahamowicz, M. Modelling time-dependent hazard ratios in relative survival: application to colon cancer. *J Clin Epidemiol* **54,** 986–996 (2001).

102. Dickman, P. W., Sloggett, A., Hills, M. & Hakulinen, T. Regression models for relative survival. *Statistics in medicine* **23,** 51–64 (2004).

103. Abrahamowicz, M. & MacKenzie, T. A. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med* **26,** 392–408 (2007).

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

105. Giorgi, R. *et al.* A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat Med* **22,** 2767–2784 (2003).

106. Nelson, C. P., Lambert, P. C., Squire, I. B. & Jones, D. R. Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med* **26,** 5486–5498 (2007).

107. Remontet, L., Bossard, N., Belot, A., Estève, J. & French network of cancer registries FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Stat Med* **26,** 2214–2228 (2007).

108. Mahboubi, A. *et al.* Flexible modeling of the effects of continuous prognostic factors in relative survival. *Stat Med* **30,** 1351–1365 (2011).

109. Quantin, C. *et al.* Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *Am. J. Epidemiol.* **150,** 1188–1200 (1999).

110. Wynant, W. & Abrahamowicz, M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in medicine* **33,** 3318–37 (2014).

111. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. & Rosati, R. A. Regression modelling strategies for improved prognostic prediction. *Stat Med* **3,** 143–152 (1984).

112. Royston, P. & Sauerbrei, W. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. (John Wiley, 2008).

113. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7,** 1–26 (1979).

114. Sauerbrei, W. & Schumacher, M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med* **11,** 2093–2109 (1992).

115. *StataCorp*. (2017).

116. Lambert, P. C. & Royston, P. Further development of flexible parametric models for survival analysis. *Stata Journal* 265–290 (2009).

117. Durrleman, S. & Simon, R. Flexible regression models with cubic splines. *Stat Med* **8,** 551–561 (1989).

118. Remontet, L., Bossard, N., Iwaz, J., Estève, J. & Belot, A. Framework and optimisation procedure for flexible parametric survival models. *Stat Med* **34,** 3376–3377 (2015).

119. Skyrud, K. D., Bray, F. & Møller, B. A comparison of relative and cause-specific survival by cancer site, age and time since diagnosis. *International journal of cancer. Journal international du cancer* (2013). doi:10.1002/ijc.28645

120. Howlader, N. *et al.* Improved estimates of cancer-specific survival rates from population-based data. *J. Natl. Cancer Inst.* **102,** 1584–1598 (2010).

121. Crocetti, E., De Lisi, V., Gafà, L., Sechi, O. & Mangone, L. Net survival: comparison between relative and cause-specific survival estimates. *Epidemiologia e prevenzione* **25,** 32–6 (2001).

122. Sarfati, D., Blakely, T. & Pearce, N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *International journal of epidemiology* **39,** 598–610 (2010).

123. Percy, C. L., Miller, B. A. & Gloeckler Ries, L. A. Effect of changes in cancer classification and the accuracy of cancer death certificates on trends in

cancer mortality. *Annals of the New York Academy of Sciences* **609,** 87–97; discussion 97–9 (1990).

124. Percy, C., Ries, L. G. & Van Holten, V. D. The accuracy of liver cancer as the underlying cause of death on death certificates. *Public health reports (Washington, D.C. : 1974)* **105,** 361–7 (1990).
125. Allemani, C. *et al.* Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet* **385,** 977–1010 (2015).
126. Crocetti, E. *et al.* Trends in net survival from breast cancer in six European Latin countries: results from the SUDCAN population-based study. *Eur. J. Cancer Prev.* **26 Trends in cancer net survival in six European Latin Countries: the SUDCAN study,** S85–S91 (2017).
127. Bossard, N. *et al.* Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM). *European journal of cancer (Oxford, England : 1990)* **43,** 149–60 (2007).
128. Roche, L. *et al.* 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* **132,** 2359–2369 (2013).
129. Dickman, P. W., Lambert, P. C., Coviello, E. & Rutherford, M. J. Estimating net survival in population-based cancer studies. *Int. J. Cancer* **133,** 519–521 (2013).
130. Seppä, K., Hakulinen, T., Läärä, E. & Pitkäniemi, J. Comparing net survival estimators of cancer patients. *Stat Med* **35,** 1866–1879 (2016).
131. Minicozzi, P. *et al.* Quality analysis of population-based information on cancer stage at diagnosis across Europe, with presentation of stage-

specific cancer survival estimates: A EUROCARE-5 study. *Eur. J. Cancer* **84,** 335–353 (2017).

132. Sant, M. *et al.* Survival of women with cancers of breast and genital organs in Europe 1999–2007: Results of the EUROCARE-5 study. *European Journal of Cancer* **51,** 2191–2205 (2015).
133. Rossi, S. *et al.* The EUROCARE-5 study on cancer survival in Europe 1999–2007: Database, quality checks and statistical analysis methods. *European Journal of Cancer* **51,** 2104–2119 (2015).
134. Mounier, M. *et al.* Changes in dynamics of excess mortality rates and net survival after diagnosis of follicular lymphoma or diffuse large B-cell lymphoma: comparison between European population-based data (EUROCARE-5). *The Lancet Haematology* **2,** e481–e491 (2015).
135. Gatta, G. *et al.* Geographical variability in survival of European children with central nervous system tumours. *European Journal of Cancer* **82,** 137–148 (2017).
136. Matz, M. *et al.* The histology of ovarian cancer: worldwide distribution and implications for international survival comparisons (CONCORD-2). *Gynecol. Oncol.* **144,** 405–413 (2017).
137. Bonaventure, A. *et al.* Worldwide comparison of survival from childhood leukaemia for 1995-2009, by subtype, age, and sex (CONCORD-2): a population-based study of individual data for 89 828 children from 198 registries in 53 countries. *Lancet Haematol* **4,** e202–e217 (2017).
138. Kvåle, R. *et al.* Prostate and breast cancer in four Nordic countries: A comparison of incidence and mortality trends across countries and age groups 1975-2013. *Int. J. Cancer* **141,** 2228–2242 (2017).

139. Cronin-Fenton, D. *et al.* Breast cancer recurrence, bone metastases, and visceral metastases in women with stage II and III breast cancer in Denmark. *Breast Cancer Res. Treat.* (2017). doi:10.1007/s10549-017-4510-3

140. Lagendijk, M. *et al.* Breast conserving therapy and mastectomy revisited: Breast cancer-specific survival and the influence of prognostic factors in 129,692 patients. *Int. J. Cancer* (2017). doi:10.1002/ijc.31034

141. Spronk, P. E. R. *et al.* Variation in use of neoadjuvant chemotherapy in patients with stage III breast cancer: Results of the Dutch national breast cancer audit. *Breast* **36,** 34–38 (2017).

142. Jinks, R. C., Royston, P. & Parmar, M. K. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Medical Research Methodology* **15,** (2015).

143. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55 (1983).