Statistical models in cancer survival - Application to study of prognostic factors in the presence of incomplete data

Luís Jorge Lopes Botelho Antunes
Programa Doutoral em Matemática Aplicada
Departamento de Matemática
2018

Orientador

Denisa Mendonça, Professora Associada,
Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto

Coorientador

Bernard Rachet, MD PhD FFPH, London School of Hygiene and Tropical Medicine

’The poor stay poor
The rich get rich
That’s how it goes
Everybody knows’
Leonard Cohen

Acknowledgments

It has been a long walk to reach this end. I crossed myself during this time with many people to which I am very grateful and to which I would like to acknowledge.

To my supervisor Professor Denisa Mendonça for all her support, dedication and long hours of healthy discussions. Without her continuous interest in my work it would have been easy to lose myself. Thank you for not having give up on me.

To my co-supervisor Professor Bernard Rachet for having accepted to supervise this work, for all the fruitful discussions and time spent correcting my texts, for his availability for all the Skype meetings, for his travels to Portugal and for receiving me in London.

Many thanks to the Director of the Department of Epidemiology of IPO-Porto, Professor Maria José Bento, for all the support during the development of this thesis, for suggesting the motivating real world research question and for her compliance with the use of data from the North Region Cancer Registry of Portugal.

A big thanks to Aurélien Belot for his selfless help in discussing the statistical and software questions that have come across this work.

My gratitude to my other co-authors in the studies developed for all their fruitful contributions, Ana Isabel Ribeiro, Camille Maringe, Hadrien Charvat and Edmund-Njeru Njagi.

To Professor Pedro Oliveira and Professor Maria Fátima Pina for their availability to discuss methodological questions related to my work.

To all my colleagues from the Department of Epidemiology, Anabela, Clara, Filipa, Roxanne, Tatiana and Vânia and to my colleagues from ORLab Andreia, Marina and Patrícia my gratitude for all their companionship and support along these last years. A special thanks to Beatriz, that unfortunately left us too soon, a big thanks for all her help and friendship.

To my PDMA colleagues for sharing common anxieties and anguishes. A special thanks to Laetitia for all her support and patient to discuss my work. We started this voyage together but she took the highway and I took the national road.

A todos os meus Amigos, vocês sabem quem são, os meus agradecimentos pela vossa contínua amizade e principalmente por se importarem.

Aos meus sogros, Jorge e Fernanda, obrigado por tudo. Não é preciso enumerar toda a ajuda que me deram e sem a qual teria sido muito mais difícil chegar ao fim desta maratona.

Ao meu pai, obrigado pela veia matemática. À minha mãe obrigado pela preocupação e pelos 'empurrões' para levar este trabalho até ao fim.

Joana, minha companheira de vida. Obrigado por todo o apoio, carinho, compreensão ao longo destes longos anos e especialmente por me manteres 'on track'.

Aos meus filhos Rita e Matias. Sem eles a minha vida seria muito mais cinzenta. Obrigado filhotes pelas alegrias que me dão e também pelas vossas traquinices.

Abstract

Cancer is a major public health issue. More than fifty thousand new cases of cancer are diagnosed in Portugal every year. The population ageing will lead to the increase of the number of newly cases in the future.

Advances in cancer diagnosis and treatment methods are conducing to improvements in survival outcomes. These gains may however not be transversal to all population. Population-based survival analysis is a fundamental tool for the evaluation of the effectiveness of cancer patient care provided to a population as well as to detect its heterogeneities.

In survival analysis, the outcome variable is time to an event, being death the event of interest in the present PhD work. In the context of population-based cancer survival analysis, cause of death is seldom available or is unreliable precluding the use of cause-specific survival. The disease-relate survival must be obtained indirectly, assuming that the observed hazard can be decomposed in two additive parcels: the excess hazard (disease related) and population hazard (other causes related). The survival directly calculated from the excess hazard is the net survival. This can be defined as the survival that would be observed if the only possible underlying cause of death was the disease under study. This measure has the advantage of being independent from background mortality and thus can be used to compare survival between subgroups with heterogeneous all-causes mortality.

The driving research question that motivated this work was the evaluation of socioeconomic inequalities in survival from cancer. Several statistical research questions arose from this motivating subject. The objectives of the studies performed were to: compare methods to estimate age-standardised net survival, using non-parametric and parametric model-based approaches; analyse and extend existing methods to model the excess hazard function in the presence of missing data on covariates; evaluate the association between socioeconomic factors and survival from cancer using net survival estimation and excess hazard modelling; sensitivity analysis of results to different assumptions on background mortality.

The main contributions of the developed work were:

- Methods to age-standardised net survival were studied and an alternative model-based approach was proposed.
- Multiple imputation methods that guarantee the compatibility between the imputation and substantive models were extended to accommodate excess hazard models.
- Deprivation-specific life tables were built for Portugal using multivariable flexible models.
- The methodology to perform evaluations of socioeconomic inequalities in survival from cancer for patients was set-up. For the first time, this evaluation was performed for patients diagnosed in the North region of Portugal.

From the studies developed, the following conclusions were draw:

- The best method to age-standardise net survival is still an open question. It has been shown that the proposed method can be a valid alternative to the conventional methods, specially in the presence of sparse data.
- The standard multiple imputation methods to handle missing data in excess hazard models with missing information on covariates can have a poor performance. The developed extension of the SMC-FCS algorithm for this context presented higher performance.
- Persistent socioeconomic inequalities in overall mortality were found for Portugal, being these larger in men than in women.
- No evidence of consistent socioeconomic inequalities in survival from colorectal cancer for patients diagnosed in the North region of Portugal were found.

Keywords: Net survival, Excess hazard, Missing data, Multiple Imputation, Age-standardisation, Population-based, Cancer, Socioeconomic inequalities, Life tables.

Resumo

As doenças oncológicas representam um importante problema de saúde pública. Mais de cinquenta mil novos casos de cancro são diagnosticados em Portugal por ano. Espera-se que o crescente envelhecimento da população leve a um aumento deste número no futuro.

Os avanços nos métodos de diagnóstico e tratamento de cancro têm conduzido a resultados mais favoráveis em termos de sobrevivência à doença. Estes ganhos podem, no entanto, não ser transversais a toda a população. A análise de sobrevivência de dados populacionais é uma ferramenta fundamental para a avaliação da eficácia dos cuidados de saúde prestados aos doentes oncológicos numa população assim como para detetar as suas heterogeneidades.

Na análise de sobrevivência, a variável resposta é o tempo até um evento, sendo a morte o evento de interesse no presente trabalho de doutoramento. No contexto da análise de sobrevivência ao cancro usando dados de base populacional, a causa da morte raramente está disponível ou não é fiável, limitando o uso da sobrevivência por causa específica. A sobrevivência relacionada com a doença é obtida indiretamente, assumindo que o risco observado pode ser decomposto em duas parcelas: o excesso de risco (relacionado com a doença) e o risco populacional (relacionado com outras causas). A sobrevivência calculada diretamente a partir do excesso de risco é a sobrevivência 'net'. Esta pode ser definida como a sobrevivência que seria observada se a única causa subjacente de morte possível fosse a doença em estudo. Esta medida tem a vantagem de ser independente da mortalidade de base e, portanto, poder ser usada para comparar a sobrevivência entre subgrupos com diferente mortalidade global.

A questão de investigação que motivou este trabalho foi a avaliação das desigualdades socioeconómicas na sobrevivência ao cancro. Diversas questões relativas à metodologia estatística surgiram do problema prático. Os objetivos dos estudos realizados foram: comparar métodos para estimar a sobrevivência 'net' padronizada para a idade, utilizando abordagens baseadas em estimadores não paramétricos e modelos paramétricos; analisar e estender os métodos existentes para modelar a função de excesso de risco na presença de dados omissos nas covariáveis; avaliação da associação entre fatores so-

cioeconómicos e a sobrevivência ao cancro utilizando estimativas de sobrevivência 'net' e modelação de excesso de risco; análise de sensibilidade dos resultados a diferentes hipóteses sobre mortalidade de base.

As principais contribuições do trabalho desenvolvido foram:

- Métodos para sobrevivência 'net' padronizada para a idade foram estudados e uma abordagem alternativa baseada em modelos foi proposta.
- Os métodos de imputação múltipla que garantem a compatibilidade entre os modelos de imputação e os modelos de análise foram adaptados para acomodar modelos de excesso de risco.
- Foram construídas para Portugal tábuas de mortalidade específicas por nível de privação usando modelos flexíveis multivariável.
- A metodologia para realizar avaliações de desigualdades socioeconómicas na sobrevivência de doentes oncológicos foi montada. Pela primeira vez, esta avaliação foi realizada para doentes diagnosticados na região Norte de Portugal.

As principais conclusões dos estudos desenvolvidos foram:

- O melhor método para padronizar a sobrevivência 'net' ainda é uma questão em aberto. Foi demonstrado que o método proposto pode ser uma alternativa válida aos métodos convencionais, especialmente na presença de pequenas amostras.
- Os métodos padrão de imputação múltipla para lidar com dados omissos em modelos de excesso de risco com informações omissa em covariáveis podem ter um desempenho fraco. A adaptação desenvolvida do algoritmo SMC-FCS para este contexto apresentou um melhor desempenho.
- Foram encontradas desigualdades socioeconómicas persistentes na mortalidade geral em Portugal, sendo as desigualdades maiores nos homens do que nas mulheres.
- Não foram encontradas evidências de desigualdades socioeconómicas consistentes na sobrevivência ao cancro colorretal em doentes diagnosticados na região Norte de Portugal.

Palavras Chave: Sobrevivência 'net', Excesso de risco, Dados omissos, Imputação múltipla, Padronização para a idade, Bases populacionais, Desigualdades socioecónomicas, Tábuas de mortalidade.

Contents

List of Tables

List of Figures

List of abbreviations

AIC Akaike Information Criterion

ASNS Age standardised net survival

ASR Age-standardised incidence rate

BIC Bayesian Information Criterion

CI Confidence Interval

EDI European Deprivation Index

EHR Excess Hazard Ratio

FCS Fully Conditional Specification

GLM Generalised linear model

ICD International Classification of Diseases

ICD-O International Classification of Diseases for Oncology

IPW Inverse probability weighting

JM Joint Model

MAR Missing at random

MCAR Missing completely at random

MCMC Markov chain Monte Carlo

MI Multiple imputation

MICE Multiple Imputation Chained Equations

MNAR Missing not at random

NS Net survival

OS Overall survival

PP Pohar-Perme

RON National Cancer Registry

RORENO North Region Cancer Registry of Portugal

SES Socioeconomic status

SMC-FCS Substantive model compatible-Fully conditional specification

Chapter 1

Introduction

1.1 Motivation

1.1.1 Survival analysis

Survival analysis is a set of statistical procedures for data analysis that is used when the outcome variable of interest is time from a well-defined time point until the occurrence of an event of interest [8]. In medical applications, the time origin corresponds usually to the recruitment of an individual to a particular study. It can be the date of diagnosis of a disease, the date of treatment initiation or for instance the beginning of follow-up after exposure to a certain factor. The final event, the event of interest, can be death, relapse, disease incidence or other specific event. The necessity of special methods to deal with this type of data is justified by the existence of censored times. A time is said to be censored when the event of interest is not observed before the end of follow-up. Another characteristic that asks for special methods is the typical asymmetric shape of survival time distributions.

Survival analysis has been used in medical research for a long time. It is possible to find in the literature a set of books dedicated to this subject [8, 9, 10, 11, 12]. Over the last decade, a growing number of research studies have been done in Portugal in the biomedical field where it is possible to find survival analysis for different applications [13, 14, 15, 16, 17].

Initially, survival methods were developed to deal with overall survival, representing the probability that an individual is still alive at time t after entering the study. The Kaplan-Meier non-parametric estimator of survival curves proposed in 1958 [18] and the Proportional Hazards Cox model proposed in 1972 [19] still belong to the current survival analysis practices of many researchers although many developments in the modelling approaches have been made meanwhile, namely, in parametric modelling.

In population-based datasets, especially those related to cancer registration, overall survival is not the most adequate measure to evaluate survival from a specific disease neither it is adequate to compare different populations. Overall survival depends not only on cancer mortality but also on other causes of mortality. It is thus necessary to have a survival measure that is independent from the background mortality. In population-based cancer survival analysis, the cause of death is seldom available or it is unreliable, turning necessary to obtain indirectly the mortality attributable to the disease. Berkson introduced the concept of relative survival for the first time in 1942 [20]. The objective was to estimate survival due to the disease ('net survival') by comparing observed survival with the survival of a similar group of individuals (relatively to demographic characteristics) taken from the general the population. Nowadays, all the methods that use this comparison are said to be in the relative survival framework.

Although there are many literature and studies in general survival analysis, the literature in the relative survival setting is more scarce. There have been many theoretical developments in the estimation of net survival in the last years. The acknowledgement in the end of the first decade of this millennium that the net survival estimators used until then were biased [21] and the introduction of the new net survival estimator by Pohar-Perme in 2012 [22] had an important impact in cancer survival analysis. Modelling of the excess hazard, quantity directly related to the net survival, has also evolved in the last years with the increasing use of flexible parametric models [23, 24, 25].

1.1.2 Cancer survival

Although the survival analysis methods in the relative survival setting can be used in other types of diseases, cancer is undoubtedly the most common application of this methodology. Cancer is the second leading cause of death worldwide and it is expected that in 2018 will be responsible for about nine and an half million deaths [26]. Its importance in the public health perspective is growing as life expectancy at birth increases and as human population is ageing, specially in the more developed countries. The impact of this group of diseases in the health of populations is substantial and demands for rigorous policies for its management and control. Population-based cancer registries are the source *par excellence* of the information needed to plan health care resources to cope with this burden. In Portugal, these registries were set up in 1988 and since then they have been providing regularly important cancer statistics for the country.

Several different measures are commonly used to describe cancer statistics, namely, incidence, prevalence, mortality and survival. While incidence is a result of risk factors acting in a population, survival is the key measure to evaluate cancer patient care. Survival statistics are of major importance to clinicians and policy makers as well as for the patients themselves. While clinical studies are usually designed to compare specific treatments in a selected group of patients, evaluating the effectiveness of cancer patient care in a certain region or country is only possible through population-based studies, since they produce results that are representative of the entire population. These studies use data collected by population-based cancer registries, which are responsible for recording all new cancer cases occurring in the covered area and for doing the respective follow-up of patients' condition.

International comparison of survival probabilities from cancer should take into account differences in patients population age structure since survival from cancer is often age dependent. This is usually achieved through direct age-standardization using a common age distribution standard. The direct age-standardization implies the estimation of survival by age group. For certain cancer sites, the extreme age groups (youngest or oldest, depending on the cancer) are sparse and their net survival estimates are either very variable or even impossible few years after diagnosis. When these situations occur, the unstandardised estimates are presented instead of the standardised survival estimates leading to incorrect comparisons of survival between different regions. The evaluation of different strategies to estimate net survival for small datasets, as is the case with rare cancers, deserves further developments.

1.1.3 Missing data

Data collected by population-based cancer registries is very often incomplete in some key explanatory variables of survival, such as stage of disease at diagnosis, morphology or grade [27]. One of the possible ways of dealing with incomplete data is to exclude from the analysis all records with missing data on any of the variables (complete-case analysis). This leads to an obvious decrease in sample size and the corresponding decrease in statistical power. It can also lead to biased results when the data are not missing completely at random, i.e. the missingness depends on the observed and/or on unobserved data [28]. Other methods involve single imputation. In these, each missing value is replaced by an estimate of it and then the completed dataset is analysed as if the imputed values were observed ones. This type of approach leads, in most situations, to invalid inferences since it does not account with the uncertainty associated with the fact that those values were not observed [29]. On the contrary, multiple imputation (MI) is an approach that considers this uncertainty by imputing multiple values to each missing observation. The MI approach has been increasingly used in published studies since it is now easily available in most statistical programme packages. MI consists of building a set of imputed datasets based on the information available, fit the substantive (analysis) model to each completed dataset and then combine the estimates from each model to obtain the final estimate [30].

Multiple imputation has been used in the context of excess hazard models using Generalised linear model (GLM) approach with Poisson error [27] or flexible parametric models [31]. More recently, Falcaro and colleagues explored multiple imputation using cumulative excess hazard models [32] and non-parametric estimation [33].

In MI the imputation and analysis steps are separate and use different models. This can lead to incompatibility issues between the imputation and substantive model arising when the associations between variables in the substantive model are not taken into account in the imputation models or when the model is itself nonlinear. Bartlett and colleagues proposed a substantive model compatible fully conditional specification methodology [34].

This methodology has been implemented in survival analysis for the traditional Cox proportional hazards model but not to the type of models mostly used in population-based survival analysis, i.e. excess hazard models.

1.1.4 Socioeconomic inequalities in cancer survival

Often, cancer survival has been shown to vary with socio-economic group. Kogevinas and colleagues in 1997 [35] and, later, Woods and colleagues [3], reviewed various studies for different tumour sites and populations (England, Scotland, Canada, United States, Australia, Norway, ...), showing evidence for survival differences between socioeconomic (SE) groups. Differences in distribution of some important prognostic factors, such as stage and treatment, are often cited as possible reasons for the socio-economic inequalities in cancer survival but insufficient for explaining all the observed differences [36].

In population-based studies, individual information on SE status is seldom. The attribution of SE condition to each patient can alternatively be done based on the patient's geographical area of provenance. Each area must be classified according to its status using an index that reflects its level of deprivation. These ecological measures can be used as a proxy of individual deprivation if the geographical areas are sufficiently small and homogeneous regarding the SE conditions [37]. Additionally, they reflect the SE conditions of the area of residence of each patient.

Different indicators of area deprivation have been in use, either simple as education or unemployment, or composite as the Carstairs or Towsend indices [38, 39]. Recently, a new ecological SE deprivation index (European Deprivation Index EDI) has been developed for several European countries (Portugal, Spain, France, Italy, England, Slovenia), based on the same methodology across all countries [40, 41]. The index was derived from country-specific census variables that are most associated with the variables of the survey European Union-Statistics on Income and Living Conditions EU-SILC [42].

Although some studies have already addressed the influence of SE status in health outcomes in Portugal [43, 44, 45, 46], SE inequalities in cancer related survival using population-based data remain to the evaluated. To assess correctly the association between the SE status of patients and their survival from cancer in the relative survival framework, the background mortality should also reflect the different SE conditions between patients. This is achievable by using deprivation-specific life tables, i.e. information on background mortality stratified by SE status. Since no deprivation-specific life tables were available for Portugal, this question needed to be addressed.

1.2 Aims of the thesis

The overall aim of the research presented in this thesis was to critically analyse existing statistical methods and propose new ones in the estimation and modelling of net survival and excess hazard in the context of population-based cancer data in situations with scarcity or lack of information, with interest in the evaluation of socioeconomic inequalities in cancer survival. Specifically, the objectives of the developed work were:

- Comparison of methods to estimate age-standardised net survival in sparse data, using non-parametric and parametric model-based approaches, proposing an alternative approach for this estimation;
- Analyse and extend existing methods to model the excess hazard function in the presence of missing data on covariates;
- Evaluation of the association between socioeconomic factors and survival from cancer using net survival estimation and excess hazard modelling; sensitivity analysis of results to different assumptions on background mortality.

Four studies have been conducted to achieve these objectives. In Study I, the age-standardised net survival estimation with sparse data is addressed. Study II deals with the estimation of net survival and excess hazard in the context of socioeconomic inequalities in cancer survival evaluation. The Study III focus on the modelling of background mortality and the construction of deprivation-specific life tables. Study IV presents the study on the missing data on covariates in excess hazard modelling applied to the evaluation of socioeconomic factors adjusting for extent of disease. They are presented as published, submitted or prepared to submit manuscripts. The first objective is addressed in Study I, the second objective in Study IV and the third objective is addressed in Studies II, III and IV.

1.3 Thesis structure

This thesis is organized in four chapters. The first constitutes the introduction, where a short summary of the background for the research questions of interest are presented. Also, the aim and objectives of the developed work are presented. In the second Chapter, the background methodology related to cancer survival analysis, socioeconomic factors, life tables construction and missing data handling is reviewed. The studies developed in this thesis are presented in Chapter three. Chapter four concludes the work with a general discussion, presentation of the limitations of the several studies and specific considerations for further research. The abstracts of the several oral and poster communications presented within the scope of this thesis are displayed in appendix. The R code developed for the main analyses described is also presented in appendix.

Chapter 2

Background

2.1 Cancer

2.1.1 What is cancer

The word cancer is used to denominate not one but more than hundred related diseases. They all have in common being originated by an uncontrolled proliferation of cells. Many cancers form solid tumours that grow-up in organs such as stomach, lung, breast, prostate, while the cancers of blood do not form solid tumours (also called liquid tumours). Cancer cells can spread into nearby tissues or can travel to distant places in the body through the blood or the lymph system, forming metastasis of the primary tumour [47].

Tumours are characterised by the location in the body where they originate (topography), by the cell type that constitute the tumour (morphology) and by their behaviour (benign, *in-situ* or malignant).

2.1.2 Cancer epidemiology

According to the International Agency for Research on Cancer (IARC) estimates, more than 17 million cancer cases (excluding non-melanoma skin cancer) will be diagnosed worldwide during 2018. In the same year, cancer will be responsible for about 9.5 million deaths. Also, it was estimated that more than 38.6 million people had a cancer less than

5 years ago and were still alive [26].

The distribution of cancer incidence around the globe is not homogeneous (Figure 2.1). The highest Age-standardised incidence rate (ASR) (world population) can be found in Oceania ($296.7/10^5$), North America ($295.6/10^5$) and Europe ($266.7/10^5$). Africa ($126.7/10^5$), Asia ($163.1/10^5$) and Latin America and the Caribeen ($181.3/10^5$), presented lowest incidence rates.

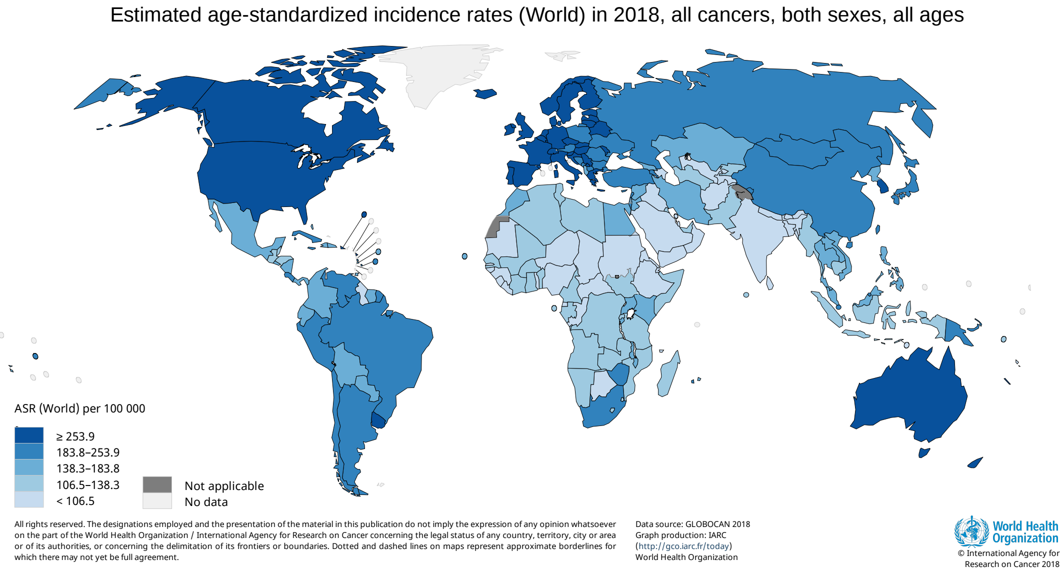
Estimated age-standardized incidence rates (World) in 2018, all cancers, both sexes, all ages
ASR (World) per 100 000
≥ 253.9
183.8–253.9
138.3–183.8
106.5–138.3
< 106.5
Not applicable
No data
All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.
Data source: GLOBOCAN 2018
Graph production: IARC
(http://gco.iarc.fr/today)
World Health Organization
World Health Organization
© International Agency for Research on Cancer 2018

Figure 2.1: Age-standardized incidence of cancer in the world (2018) [6].

In Portugal, 46724 new cases were diagnosed in 2010, corresponding to an ASR (world population) of 238.8/100 000 [48]. The most common were prostate, colorectal and lung cancers, in men, and breast, colorectal and thyroid cancers, in women.

The aetiology of cancer is not fully understood. There are however some known risk factors that affect the chances of developing cancer. The most important human carcinogens include tobacco, asbestos, aflatoxins and ultraviolet light. Almost 20% of cancers are associated with chronic infections (HBV, HCV, HPV, *Helicobacter pylori*). There is increasing recognition of the causative role of lifestyle factors, including diet, physical activity, and alcohol consumption. Genetic susceptibility may significantly alter the risk from environmental exposures. For most type of cancers, the risk of developing the disease increases with age [49].

While risk factors relate to the chance of developing cancer, prognostic factors influence the chances of surviving the disease. Prognostic factors can be divided in tumour-, patient- and environment-related [50]. The first type includes the major prognostic factor that is stage of disease at diagnosis, i.e. if the disease is still at an initial point of its natural development or it is already in a very advanced stage. Histological grade can also be included in the first group. The patient-related factors include demographic characteristics as age, sex and ethnicity. Also the performance status, comorbidities or immune status of the patient can influence their survival. The environmental factors are those that are external to the patient. Choice and quality of treatment, access to care, health-care policy are among this group of factors.

2.1.3 Cancer Registries

Cancer registries are organizations for the systematic collection, storage, analysis, interpretation and reporting of data on subjects with cancer [51]. Registries can be hospital- or population-based. The first type of registry is responsible for collecting information from a single institution. Data can be used for reviewing clinical performance but can not be used to produce measures of cancer burden in a population since its catchment area is not precisely defined. On the other hand, population-based registries collect information from a well-defined population. These registries can thus produce statistics of incidence or survival representative of its catchment area. The information produced by population-based registries can and should be used for monitoring and assessing the effectiveness of cancer control activities.

The two main cancer burden measures provided by the population-based cancer registries are incidence rates and survival probabilities. Incidence represents the number of new cases occurring in a certain well-defined population by period of time and by inhabitants at risk of developing the disease. The analysis of incidence figures and trends allows to evaluate the impact of measures of primary prevention, i.e. measures aiming at preventing the development of the disease. Survival measures the outcome of patients already diagnosed with the disease. It thus allow to evaluate the performance of a healthcare system and the impact of healthcare policies regarding introduction of new diagnostic procedures, new therapies, innovative drugs, among others. Other cancer burden indicator of major importance is mortality. Information on this indicator is normally the responsibility of the National Statistics Offices and not of the cancer registries, as is the case in Portugal [52].

2.1.4 Classification of cancer

Cancer registries use international classification systems in order to allow homogeneity in registry procedures and enhance the comparability between different regions. The World Health Organization (WHO) defined the International Classification of Diseases for Oncology (ICD-O) [53] to code the topography (site of primary tumour) and morphology (histological type) of the tumours. The fifth digit in the ICD-O morphology codes describes the behaviour of the tumour: benign, borderline, *in situ*, malignant. Many cancer registries use the International Classification of Diseases (ICD) to present their results, namely, its 10th edition [54].

2.1.5 Cancer registration in Portugal

The North Region Cancer Registry of Portugal (RORENO) is a population-based cancer registry. It is one of the four population-based cancer registries established in Portugal that together cover the complete area of the country (mainland and the archipelagos of Azores and Madeira). The catchment area of RORENO corresponds to the North Region of Portugal. Until 2009, this area corresponded to the districts of Braga, Bragança, Porto, Viana do Castelo and Vila Real, with approximately 3.2 million inhabitants (around 30% of the Portuguese population). From 2010 onwards, the catchment area corresponds to the NUTSII North region ($\approx$ 3.6 million inhabitants). It was established in 1988, and it is responsible for collecting all new cancer cases occurring in the covered area. It is based in the Portuguese Oncology Institute of Porto (IPO-Porto). The main public hospitals register their cases directly in the RORENO database through a web-based platform. Private hospitals and pathology laboratories report their cases through spreadsheet files and are integrated in the database by the staff in the cancer registry. All registered cases are controlled for its quality using the IARC rules [55]. RORENO has participated in several international studies including the European survival study EUROCARE-5 [56] and the international CONCORD-2 and CONCORD-3 studies [57, 58]. RORENO publishes incidence and survival reports on a regular basis. They are available on-line to the general public (`www.roreno.com.pt`). RORENO publishes its results using the ICD-10 classification system.

In 2017, a law (53/2017) was published creating the National Cancer Registry (RON) (*Registo Oncológico Nacional*) [59]. According to this law, the RON is a centralised registry based on a single electronic platform, with the purpose of collecting and analysing data of all cancer patients diagnosed and/or treated in mainland Portugal and in the autonomous regions, allowing the monitoring of the activity performed by the institutions, the effectiveness of organized screening and therapeutic effectiveness, epidemiological surveillance, research and to monitor the effectiveness of drugs and medical devices.

2.2 Some concepts of survival analysis

2.2.1 Introduction

Survival analysis includes a set of statistical methods developed for the analysis of survival data. The variable of interest is commonly designated as *survival time*. This represents the time from a specific time point and the occurrence of an event of interest. In cancer survival analysis, typical examples are time from diagnosis to death due to cancer or time from date of treatment until relapse or progression. Time-to-event analysis is not an exclusive of health related problems. In engineering applications, for instance, it can be used to analyse time from operation start until machine failure.

The necessity of specific methods to analyse survival data comes from the fact that commonly the event of interest does not occur for all individuals during the follow-up period (censored data). For these cases, the actual survival time is unknown. Additionally, survival times are usually right skewed and so not normally distributed [60].

The main interests in survival analysis are the study of the probability of occurring the event of interest and of the rate of its occurrence. Comparison of these quantities between different groups of individuals or finding variables that can explain them is among the objectives of a typical survival analysis [8].

2.2.2 Censoring

When the event of interest is not observed for an individual during the duration of the study, the corresponding time is said to be right censored. This type of censoring can occur for different reasons: i) the individual does not experience the event by the time of study end (administrative censoring); ii) the individual is lost to follow-up, i.e. the individual status is only known until a certain date prior to the study end; iii) the individual experiences a competing event that precludes the occurrence of the event of interest [60]. Other types of censoring can occur (left or interval). Left censoring occurs when the actual survival time is lower than the observed. If the first time the patient is observed after entering the study the event already occurred. Then, it would only be known an upper bound of the survival time but not its exact value. In interval-censoring the event is known to have occurred within an interval of time. For example, a cancer relapse occurs between two medical appointments and the exact date of the relapse is not known [10]. Only right censoring is dealt with in this thesis and for the sake of simplicity the term censoring will be used to refer to right censoring, unless explicitly mentioned otherwise.
Censoring can be informative or non-informative. If the actual survival time (t) of an individual is independent of the mechanism that leads that time to be censored in a specific time c ($c < t$), censoring is said to be non-informative. This means that the individual censored must be representative of all individuals, sharing the same values of the prognostic variables, that survived until that censoring time. On the other hand, if censoring time depends on the probability of the event occurrence then censoring is informative. Most of the methods used in survival analysis rely on the assumption of non-informative censoring. Administrative censoring can usually be considered non-informative since all patients still alive are censored at the same time, independently of factors that can influence their survival.

2.2.3 Survival function and hazard function

In survival analysis there are two functions of major interest, namely, the survival function and the hazard function. Their definition, the relationship between them and between other functions of interest and important mathematical notation are described next.

Let T be a non-negative random variable, absolutely continuous, that represents the survival time. Considering that the probability distribution of this variable is described by the density function $f(t)$, the probability of T being lower or equal than a specific time t is given by the distribution function $F(t)$ (2.1).

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{2.1}$$

The survival function is defined as being the probability of T being greater than a specific time t and is the complement of the distribution function (2.2).

$$S(t) = P(T > t) = 1 - F(t), \quad t \geq 0. \tag{2.2}$$

The survival function is a continuous monotonically non-increasing function. Also, $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$.

Another quantity of interest is the hazard function. This function represents the instantaneous event rate conditioned on having survived until time t (2.3).

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.3}$$

The hazard function is a non-negative function but, as opposed to the survival function, it is not necessarily monotonous. Also, it has no upper-bound. This function gives information on the evolution with time of the instantaneous event rate while the survival function reflects the cumulative non-occurrence of the events. Both functions are mathematically related (Equations: 2.4, 2.5):

$$S(t) = exp\left(-\int_0^t \lambda(u)du\right), \tag{2.4}$$

and,

$$\lambda(t) = -\frac{1}{S(t)}\frac{dS(t)}{dt}. \tag{2.5}$$

Also, the density, distribution and hazard functions are related by:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}. \tag{2.6}$$

The cumulative hazard is a non-negative and monotonically increasing function (Equation 2.7).

$$\Lambda(t) = \int_0^t \lambda(u)du, \quad t \geq 0. \tag{2.7}$$

The survival function can be directly obtained from the cumulative hazard function (Equation 2.8). Higher cumulative hazards imply lower survival.

$$S(t) = exp(-\Lambda(t)). \tag{2.8}$$

2.2.4 Conditional survival

In clinical applications, the survival function can sometimes be not very informative for patients who have already survived a certain amount of time, since it is based on all patients, including the ones with very bad prognosis that have very short survival times. The probability of surviving a certain additional amount of time conditioned on the fact that the patient has already survived some time (t_i, $i > 0$), can be more informative (2.9) and is often presented in cancer survival reports.

$$S_{t|t_i} = P(T > t|T > t_i) = exp\left(-\int_{t_i}^t h(u)du\right) = \frac{S(t)}{S(t_i)} \tag{2.9}$$

2.2.5 Measures of survival

Cancer patients may die from the disease itself or they may die from other causes. These two are competing events since one person that dies from other causes can no longer die from cancer. It is assumed that the observed hazard (λ_{Oi}) of a particular individual can be decomposed in two additive components: the hazard due to the disease in question (λ_{Ei}) and the hazard due to other causes (λ_{Pi}) (2.10).

$$\lambda_{Oi}(t) = \lambda_{Ei}(t) + \lambda_{Pi}(t) \tag{2.10}$$

This decomposition holds if the time to death from the disease and the time to death from other causes are conditionally independent given a known set of covariates [22]. In the usually designated relative survival setting, it is assumed that the hazard due to other causes is given by population mortality tables. It is also assumed that the disease-specific mortality included in the overall mortality is negligible.

In the context of cancer survival analysis, there are several different questions one might be interested in. For example, estimating the survival of a cohort of patients, independently of the cause of death; knowing the proportion of patients that died from cancer and the proportion that died from other causes; comparing the survival of a cohort of patients with the survival of a similar group of individuals but cancer-free; comparing healthcare system's performance between countries. The measure of survival for each situation must be chosen according to the objective of the analysis. Four most common measures are presented below [61], giving special emphasis to net survival which is the main measure used along this thesis.

Overall survival

Overall survival, also designated as all-cause survival or observed survival, is defined as the probability that a individual is still alive after a certain time t:

$$S_O(t) = P(T > t) = exp\left(-\int_0^t \lambda_O(u)du\right). \tag{2.11}$$

In this situation, one is not interested in the cause of death of each patient so the survival is related with the overall hazard rate of death (λ_O).

Cause-specific crude mortality

The cause-specific crude mortality gives the probability of dying from one particular cause up to a time t. It splits the overall mortality ($1-$overall survival) into the probability of dying from the specific cause (F_E) and the probability of dying up to t from other causes (F_P): $1 - S_O(t) = F_E(t) + F_P(t)$. The also called cumulative cause-specific mortality (F_E) is given by:

$$P(T \leq t, cause = cancer) = F_E(t) = \int_0^t S_O(u-)\lambda_E(u)du, \tag{2.12}$$

meaning that for a patient to die from a specific cause at time t they must have survived from all causes just until the time immediately before.

Relative survival ratio

Relative survival ratio ($S_R(t)$) compares the observed survival in a group of patients to the survival that group would experience if they were free of the disease. This second quantity is called the expected survival ($S_P(t)$) and it is assumed that it can be obtained from general population life tables. The comparison is made by calculating the quotient between the two survival:

$$S_R(t) = \frac{S_O(t)}{S_P(t)}. \tag{2.13}$$

This means that if the relative survival ratio is equal to 1, the survival of the group of patients under study is equal to the expected survival that a group of individuals free of the specific disease under study but with the same demographic characteristics would have.

Net survival

Net survival is obtained from the hazard function related with the disease ($\lambda_E(t)$):

$$S_N(t) = exp\left(-\int_0^t \lambda_E(u)du\right). \tag{2.14}$$

It represents the survival that would be observed in the hypothetical situation where the disease under study would be the only possible cause of death [22]. This concept is further explored below.

2.2.6 Non-parametric estimation

To estimate survival for a group of patients it is necessary to have information not only on survival times but also on an indicator variable (2.15) that allows to distinguish each survival time as time to event t_i (considering only one event of interest) or time to censoring c_i.

$$\delta_i = \begin{cases} 1 & \text{if } t_i \leq c_i \\ 0 & \text{if } t_i > c_i \end{cases} \tag{2.15}$$

The non-parametric estimation methods of survival do not assume any specific dependence between survival and time neither between survival and any prognostic factors. It allows a first description of the data to be analysed but it does not allow adjusting for potential confounding factors.

The method of estimation depends on the measure of survival being estimated. The four most common measures described above are considered again.

Overall survival

Overall survival is estimated using equation (2.11). It is thus necessary to estimate the cumulative hazard increment $\lambda_O(u)du$ which can be given by:

$$\hat{\lambda}_O(u)du = \frac{\textit{number of events on a short interval of time}}{\textit{number of individuals at risk}}. \tag{2.16}$$

Two possible non-parametric estimators of overall survival are the Kaplan-Meier and the Nelson-Aalen estimators.

- Kaplan-Meier estimator

  The most widely used estimator of the survival function is the Kaplan-Meier non-parametric estimator also called product-limit [18]. Let $t_1 < t_2 < \cdots < t_k$ be distinct times at which k events occur. Since times of each event are assumed to be independent from each other, the cumulative survival probability can be obtained by the product of the probabilities of surviving from one interval to the other:

$$\hat{S}_O(t_j) = \hat{S}_O(t_{j-1}) \cdot \hat{P}(T > t_j | T \geq T_j) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i}\right),$$

where n_j represents the number of individuals at risk just before t_j and d_j the number of events at t_j. This estimator assumes that the survival function is constant between any two consecutive events. Each time an event occurs the estimate is updated. Censored individuals in each interval reduce the number of individuals at risk of having the event and contrarily to the events affect only the denominator in the fraction d_j/n_j.

Variance estimation

The variance of the survival estimate obtained by the Kaplan-Meier estimator can be given by the Greenwood formula (2.17). This assumes that the number of individuals who survive each time interval follows a binomial distribution with parameters n_j and p_j (true probability of survival) and approximates the exact value through a first-order Taylor series approximation.

$$\hat{\sigma}^2_{S_O}(t) \approx \hat{S}^2_O(t) \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \tag{2.17}$$

- Nelson-Aalen estimator

  The cumulative hazard function can be estimated non-parametrically using the Nelson-Aalen estimator (2.18) [62, 63].

$$\hat{\Lambda}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j} \tag{2.18}$$

This represents the cumulative sum of the estimated probabilities of the event occurrence, where d_j and n_j have the same meaning as above. The survival is then obtained by:

$$\hat{S}_O(t) = exp\left(-\sum_{t_j \leq t} \frac{d_j}{n_j}\right) \tag{2.19}$$

Variance estimation

The variance of the Nelson-Aalen estimator can be given by [64]:

$$\hat{\sigma}^2(t) = \sum_{t_j \leq t} \frac{(n_j - d_j)d_j}{(n_j - 1)n_j^2} \tag{2.20}$$

Cause-specific crude mortality

To estimate the cause-specific crude mortality (2.12) it is necessary to estimate the observed survival and the hazard rate related to the disease of interest. The first quantity can be estimated using the Kaplan-Meier estimator described above and the second by doing the difference between the observed hazard and the population hazard. Further details on the estimation method can be found in the literature [65].

Relative survival ratio

The relative survival ratio is calculated as the ratio between the observed survival and the expected survival. The observed survival can be estimated using the Kaplan-Meier estimator described above. The expected survival is obtained from general population life tables, calculated from mortality and population data, usually available at the national statistics offices. The Human Mortality Database (available at `http://www.mortality.org`) provides life tables for a considerable number of countries including Portugal. These life tables are at least stratified by sex, age and calendar year. For some countries or regions it is also possible to obtain life tables stratified by socioeconomic condition, ethnicity or other demographic variables. It is assumed that the life tables represent the mortality of individuals free of the disease in study which is an acceptable assumption when the disease in study is relatively rare [66].

Three different estimators of the expected survival can be found in the literature. These estimators are usually referred as Ederer I [66], Ederer II [67] and Hakulinen [68]. They differ regarding the time each individual is considered to be at risk.

Ederer I

In this method, the general population matched individuals are considered indefinitely at risk. The time at which the correspondent patient dies or is censored does not affect the expect survival.

Ederer II

In Ederer II estimator, the matched individuals are at risk while the corresponding individual from the study population is not censored or dies.

Hakulinen

In the Hakulinen estimator the survival time of the matched individual is censored at the same time as the patient's survival time but if the patient dies, the matched individual is at risk until the end of the study. Information about the date of the end of follow-up is thus necessary in this method.

Net survival

Net survival is described in more detail in the following section since it was the main measure used over this study.

2.3 Net survival

2.3.1 Introduction

In the analysis of cancer survival data, the interest usually lies on analysing time since disease diagnosis until death. Since cancer patients can died not only from cancer but also from other causes, when comparing cancer survival between different periods of diagnosis, different regions or different socioeconomic groups for instance, it is important to have a measure that is independent from background mortality. Overall survival is thus not adequate for this type of comparison, especially in elderly patients for which other cause mortality is higher. Cause-specific survival, where only death caused by the disease in question is considered an event and all others are censored, depends on the knowledge of cause of death for all patients. In population-based data sets, this information is usually not available or is not reliable. Crude mortality quantifies the actual contribution of the cancer to overall mortality. However, it is not good for comparing different regions since it also affected by background mortality [61].

Net survival is defined as the survival that would be observed in the hypothetical situation that the disease is the only cause of death possible. Although this survival is not observable in the real world, it is of interest. It is the only measure that allows comparisons between different populations (originated from different regions, calendar years or other factors) since it is independent of other causes mortality [22, 61].

2.3.2 Estimation

Net survival for an individual i is given by the integral of the excess hazard function, i.e. the hazard due to the specific disease in study (2.21).

$$S_{N_i}(t) = exp\left(-\int_0^t \lambda_{E_i}(u)du\right) = exp\left(-\Lambda_{E_i}\right) \tag{2.21}$$

The net survival of the overall cohort of n patients is given by the average of the individual survivals:

$$S_N(t) = \frac{1}{n}\sum_{i=1}^{n} S_{N_i}(t) \tag{2.22}$$

Pohar-Perme and colleagues proposed in 2012 [22] a new estimator (PP) for net survival. The hazard due to cancer is given by the difference between the observed and the expected (population) hazard. However, to compensate the early drop off from the sample of the patients with higher background mortality the counting and the at-risk process of each individual are weighted using his/her expected survival time distribution.

Let $N_i(t) = I(T_i \leq t, T_i \leq C_i)$ and $Y_i(t) = I(T_i \geq t, C_i \geq t)$ denote the counting process and the at-risk process for each individual in the sample, where T_i denotes the time to death from any cause and C_i the time to censoring. The PP estimator weights these two processes using the inverse of the population survival probability: $N_i^w(t) = N_i(t)/S_{Pi}(t)$ and $Y_i^w(t) = Y_i(t)/S_{Pi}(t)$. The cumulative excess hazard is thus given by:

$$\hat{\Lambda}_E(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i^w(u) d\Lambda_{P_i}(u)}{Y^w(u)} \tag{2.23}$$

The variance of this estimator is given by [22]:

$$\hat{\sigma}^2(t) = \int_0^t \frac{J(u) \sum_{i=1}^n dN_i(u)/S_{Pi}^2(u)}{\left\{ \sum_{i=1}^n Y_i(u)/S_{Pi}(u) \right\}^2}, \tag{2.24}$$

where $J(t) = I\{Y(t) > 0\}$.

2.3.3 Net Survival vs Relative Survival Ratio

Relative survival ratio was used in the past as the main measure reported by cancer registries and by survival international comparison studies as it was thought to be the same as net survival. Net survival can be interpreted as the average of the ratio of overall and population survival (2.25).

$$S_N(t) = \frac{1}{n} \sum_{i=1}^n \frac{S_{Oi}(t)}{S_{Pi}(t)} \tag{2.25}$$

The relative survival ratio, however, can be written as the ratio of the average of overall survival by the average of population survival (2.26), which is different from the quantity presented above.

$$S_R(t) = \frac{\frac{1}{n}\sum_{i=1}^n S_{Oi}(t)}{\frac{1}{n}\sum_{i=1}^n S_{Pi}(t)} \tag{2.26}$$

Older international studies such as the first edition of the CONCORD study [69] or the first editions of the European study EUROCARE [70] have used the relative survival ratio estimated using the Hakulinen method. The author himself, however, noticed that this method produced biased estimates and inconsistencies and recommended to use other available methods [21]. A couple of years later [71], Hakulinen and colleagues proposed that the method designated by Ederer II, initially developed in the beginning of the sixties of the twentieth century, should be adopted. This method was later used in EUROCARE-5 [72]. Maja Pohar-Perme proposed her new estimator of net survival in 2012 [22]. She has proven that her estimator was the only one that estimated in fact net survival and the others estimated different quantities. Since the introduction of this new estimator, several articles have been published comparing the performance of this estimator with the performance of the traditional relative survival ratio estimators. Danieli and colleagues [73] concluded in a simulation study comparing several approaches that the PP estimator and a multivariable modelling approach were the only methods that had a good performance in all tested scenarios. Later, Roche and colleagues [74] compared the estimates obtained by the several traditional methods with the PP estimates using real cancer data. They concluded with the recommendation that the classical methods should be abandoned and the PP estimator should be adopted by cancer registries. This work deserved some criticism from other authors [75] claiming that the benefits of the new estimator were overestimated. In 2015, Lambert and colleagues [76] compared several methods for estimating age-standardised net survival, concluding that the PP estimator does not present a considerable advantage over the age-standardised Ederer II. Additionally, they considered this last estimator to have improved precision. Seppa and colleagues [77, 78] also concluded that the differences between the several approaches are small and that the PP estimates are prone to random variation for long-term follow-ups. Pohar-Perme argues that this large variability is a reflection of the lack of information available in the data and it is a characteristic of the measure and not of the estimator [61].

Nevertheless, despite these different opinions, the PP estimator is the only recognized consistent estimator of net survival besides an adequate multivariable model. It is possible to find in the literature an increasing number of publications using this methodology [79, 80, 81, 82, 83] including the largest international survival comparison studies - CONCORD-2 and CONCORD-3 [57, 58] and the SUDCAN study [84]. In this thesis, it has been used as the non-parametric estimator of net survival.

2.3.4 Comparison of net survival distributions

Until recently the comparison of net survival distributions over a given period of time was only possible using modelling approaches. It was possible to compare point estimates using a Z-test but not a net survival distribution estimated non-parametrically. Graffeo and colleagues proposed in 2016 a log-rank type test to overcome this limitation [85]. It allows the comparison of net survival distributions between two or more groups over a certain follow-up period. The test is similar to the well known log-rank test widely used for the comparison of observed survival distributions [10].
The null hypothesis of this log-rank type test is:

$$(H_0) : \forall t \in [0, T], \, \Lambda_{E,1}(t) = \cdots = \Lambda_{E,k}(t), \tag{2.27}$$

where $k \geq 2$ is the number of groups to be tested. The test compares the k cumulative hazard distributions with the expected distribution under the null hypothesis. These are estimated by the Pohar-Perme estimator. The test lies on the following assumptions:

- $(T_{P_{h,i}}, T_{E_{h,i}}, C_{h,i}, \mathbf{X}_{h,i})_{h,i}$ are mutually independent;
- $(T_{P_{h,i}}, T_{E_{h,i}}, C_{h,i}, \mathbf{X}_{h,i})_{i}$ have the same distribution;
- $T_{E_{h,i}}$ and $T_{P_{h,i}}$ are conditionally independent given $\mathbf{X}_{h,i}$;
- censoring times $C_{h,i}$ are independent of pair $T_{h,i}, \mathbf{X}_{h,i}$.

The variable T represents time to death, the indices P and E represent, respectively, the population and excess hazards, C the censoring time, $\boldsymbol{X}$ a vector of covariates and h the index of the groups in comparison. Under these assumptions, it can be shown [85] that under the null the test statistic follows approximately a chi-square distribution with $k-1$ degrees of freedom.
In order to take into account the effect of confounding covariates, a stratified version of the test was also proposed [85]. This version relaxes the assumption of independence between T_E and $\boldsymbol{X}$. It is assumed homogeneity within each stratum but heterogeneity is

allowed between strata. The null hypothesis is in this situation:

$$(H_0) : \forall t \in [0,T], \forall s \in \|1; m\| \Lambda_{E,1,s}(t) = \cdots = \Lambda_{E,k,s}(t), \tag{2.28}$$

where m is the number of strata. The test statistic continues to have an asymptotic chi-square distribution with $(k-1)$ degrees of freedom. According to the authors, the decision of using or not the stratified version should be based on epidemiological considerations [85].

2.4 Hazard modelling

2.4.1 Survival models

In the previous section the survival function was discussed and methods to estimate it non-parametrically were presented as well as a test to compare survival between groups. This type of analysis is important but it has several limitations. The estimation of survival curves allows a simple analysis of the influence of one variable in survival but it is much harder if one is interested in analysing simultaneously the possible confounding effect of other variables or interaction between them. Also, the test presented to evaluate survival differences between groups offers no estimate of the actual effect size of the variable in analysis. The use of statistical models complements the analysis presented before by allowing the simultaneous investigation of different covariates and by quantifying the effect of each variable while adjusting for the other covariates effects [86]. In this section, the most common regression models used in the context of survival analysis, and more precisely in the context of population-based cancer survival analysis, are discussed.

It is possible to have different outcome variables in the regression models: survival time, cumulative hazard or hazard. The most common family of models for survival time is the Accelerated Failure Time (AFT) models. This type of models are however rarely seen in cancer population-based studies. Much more common are models for the hazard and the cumulative hazard and these are described below.

2.4.2 Hazard models

In the relative survival setting it is assumed that the observed hazard can be split in two additive parts [87]:

$$\lambda_O(t) = \lambda_P(t) + \lambda_E(t) \tag{2.29}$$

As explained before, it is assumed that the other causes hazard can be correctly estimated from the population hazard. The $\lambda_E(t)$ represents thus the excess hazard attributable to the disease in study. Since $\lambda_P(t)$ is considered as a known quantity and the interest lies on the hazard disease related, the excess hazard is the outcome variable modelled in the regression models. The available models for modelling hazard or the excess hazard are basically the same. If we assumed λ_P to be zero, the two would be the same quantity.

The effect of the covariates in the excess hazard can be considered additive (2.30) or multiplicative (2.31). In these two formulations $\lambda_0(t)$ represents the baseline excess hazard, i.e. the excess hazard function when all covariates assume the value zero (reference categories).

$$\lambda_E(t, \mathbf{x}) = \lambda_0(t) + \sum_{j=1}^{p} \beta_j x_j \tag{2.30}$$

In the first option, the effect of the covariates is supposed additive. For a given covariate, its effect is always the same independently of the value of the baseline.

$$\lambda_E(t, \mathbf{x}) = \lambda_0(t) \cdot exp\left\{ \sum_{j=1}^{p} \beta_j x_j \right\} \tag{2.31}$$

In the second formulation, the effect of the covariates is supposed to act multiplicatively on the baseline. Thus, the effect of each covariate depends on the baseline hazard level. This type of models is the most common and it is the only one mentioned from now on.

Semi-parametric models

A semi-parametric model with multiplicative covariate effects for the excess hazard can be written as [88]:

$$\lambda_E(t, \mathbf{x}) = \underbrace{\lambda_0(t)}_{\text{non-parametric}} \cdot \underbrace{exp\left\{\sum_{j=1}^{p} \beta_j x_j\right\}}_{\text{parametric}} \tag{2.32}$$

In this type of model, using a partial likelihood function in the estimation process, only the effects of the covariates are estimated leaving the excess hazard baseline unspecified. This type of model for the hazard function was introduced by Cox in 1972 [19] and is still one of the most used in the medical literature.

Pohar-Perme proposed in 2009 a new approach for fitting the model (2.32) that makes no assumptions about the form of the baseline excess hazard and is based on an expectation-maximization (EM) algorithm with the cause of death treated as missing data [89]. Since this approach does not make assumptions about the shape of excess hazard baseline, it can be used to check informally the goodness of fit of a parametric model.

Parametric models

In parametric models the excess hazard function, baseline and covariates effects, are fully specified. The baseline can be modelled using defined distribution functions such as exponential, Weibull, Gompertz, log-normal, among others. These types of models make strong assumptions about the shape of the baseline excess hazard function. They are either constant or monotonically increasing or decreasing showing low flexibility to capture the shapes of the functions found in real clinical datasets.

Estève and colleagues [87] proposed a simple model where the baseline was considered to be a piecewise constant function.

$$\lambda_E(t, \mathbf{x}) = \sum_{k=1}^{r} \tau_k I_k(t) \cdot exp\left\{\sum_{j=1}^{p} \beta_j x_j\right\} \tag{2.33}$$

Besides assuming that the excess hazard function is constant in each time period predefined, it also assumes proportional hazards, i.e. that the effect of each covariate is independent from time, and assumes the effect of each covariate to be log-linear. Higher or lower flexibility to capture the baseline shape can be achieved by increasing or decreasing the number of time periods considered. The specification of the likelihood function is simple since it is possible to obtain analytically the integral of this function.

This model has been extended successively to increase its flexibility. Instead of a piecewise constant function for the baseline, regression splines were introduced [90] that allowed to get a more realistic baseline function. Also, time-dependent effects of the covariates were introduced. Later, Remontet and colleagues [23] further extended this model to incorporate non-linear effects of covariates. Recently, Charvat and colleagues [25] proposed a model (2.34) including random effects in order to take into account the possible hierarchical structure of the data.

$$log(\lambda_E(t, \mathbf{x})) = log(f(t)) + \beta(t)\mathbf{x} + g(\mathbf{x}) + w_i, \tag{2.34}$$

where w_i is a random effect and f and β are B-Splines and g can be a linear or non-linear function of the covariates.

B-splines are a type of regression splines, i.e. piecewise polynomial functions. The pieces join at points referred to as knots and the greater the number of knots, the higher the flexibility of the function. Considering a B-spline function of order q, this function and its first $(q-2)$ derivatives are continuous at the knots. Considering a total of m interior knots $(t_1, \cdots, t_m)$ plus two boundary knots (t_0 and t_{m+1}) and plus $2(q-1)$ additional boundary knots such that $t_{-(q-1)} = \cdots = t_{-1} = t_0$ and $t_{m+1} = t_{m+2} = \cdots = t_{m+q}$. The basis functions $B_{-(q-1),q}(t), \cdots, B_{m,q}(t)$ are recursively defined by:

$$B_{j,q}(t) = \frac{t - t_j}{t_{j+q-1} - t_j} B_{j,q-1}(t) + \frac{t_{j+q} - t}{t_{j+q} - t_{j+1}} B_{j+1,q-1}(t), \tag{2.35}$$

where $j = -(q-1), \cdots, m$, $B_{j,1}(t) = 1$ if $t \in [t_j, t_{j+1}[$ and $B_{j,1}(t) = 0$ otherwise [91]. The resulting B-spline function is a linear combination of the basis functions:

$$BS(t) = \sum_{j=-(q-1)}^{m} \alpha_j B_{j,q}(t), \qquad t \in (t_0, t_{m-1}). \tag{2.36}$$

Crowther and Lambert [92] proposed a similar model for the excess hazard, also considering a flexible parametric function for modelling the baseline and time-dependent effects of the covariates:

$$log(\lambda_{Ei}(t)) = s(log(t)|\gamma_0, \mathbf{k}_0) + \mathbf{X}_i\beta + \sum_{p=1}^{P} x_{ip} s(log(t)|\gamma_p, \mathbf{k}_p) \tag{2.37}$$

In this formulation, restricted cubic splines were proposed. This type of splines imposes the constraint that the fitted function is linear beyond the boundary knots. A restricted cubic spline ($s(u|\gamma, \mathbf{k}_0)$) where γ is the parameter vector, $\mathbf{k}_0$) is the knot vector and u is the variable of interest (eg, $u = log(t)$), is defined by:

$$s(u|\boldsymbol{\gamma}, \mathbf{k}_0) = \gamma_0 + \gamma_1 s_1 + \gamma_2 s_2 + \cdots + \gamma_{m+1} s_{m+1} \tag{2.38}$$

The derived variables s_j or basis functions are given by:

$$\begin{aligned} s_1 &= u \\ s_j &= (u-k_j)^3_+ - \lambda_j (u-k_{min})^3_+ - (1-\lambda_j)(u-k_{max})^3_+ \quad \text{if } j = 2, \cdots, m+1 \end{aligned} \tag{2.39}$$

where $(u-k_j)^3_+$ is equal to $(u-k_j)^3$ if the value is positive and zero otherwise, and

$$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}} \tag{2.40}$$

Other options for a flexible modelling of the excess hazard baseline can be found in the literature such as the use of fractional polynomials [93].

Estimation

Hakulinen and Tenkanen [94] and Dickman and colleagues [95] proposed excess hazard models in the framework of generalized linear models. The first considered a binomial error structure and the second a Poisson error structure. In both approaches the model is estimated from subject-band observations, splitting the data in pre-determined time intervals and by combination of covariate values.

With the increasing computational power available nowadays the methods using individual data are preferred. Model parameters are estimated using maximum likelihood methods. When fitting a model to a set of data, one wants to estimate the set of coefficients (*β*) that maximizes the likelihood function that is the same to say that maximizes the log-likelihood function ($LogL$):

$$LogL = \sum_{i=1}^{n} logL_i \tag{2.41}$$

Considering a general excess hazard model of the form $\lambda_O = \lambda_P + \lambda_E$, the log-likelihood contribution of the i^{th} patient for the overall log-likelihood function is (ignoring the terms that do not depend on *β*):

$$logL_i(\beta|t_i, \delta_i, x_i) = \delta_i \cdot log\left[\lambda_P(a_i + t_i, y_i + t_i|D) + \lambda_E(t_i, x_i|\beta)\right] - \int_0^{t_i} \lambda_E(u, x_i|\beta)du, \tag{2.42}$$

where δ_i is the censoring indicator, a is the age at diagnosis, y is the year of diagnosis and D is the set of demographic variables.

In the general case the integral of the excess hazard function does not have a closed-form. It requires numerical integration techniques to compute it. One possible algorithm, highly efficient and implemented in several software packages, is the Gauss-Legendre quadrature of order n approximation. It transforms the integral in a weighted sum of the function to integrate evaluated at a set of n points called nodes (t_k).

$$\int_0^{t} \lambda(u)du \approx \sum_{k=1}^{n} w_k \cdot \lambda(t_k) \tag{2.43}$$

The knots position and weights do not depend on the integrand. This approximation gives the exact integral for any polynomial of degree $\leq 2n-1$.
Crowther and Lambert in their proposed model using restricted cubic splines for the baseline, proposed a refinement in the estimation procedure to improve computation time. Since beyond the boundary knots the log excess hazard is forced to be a linear function of the log time, they solve these part of the function analytically and leave the numerical integration for the interval within the boundary knots. They have shown a reduction on the number of nodes needed to obtain the same precision [96].

2.4.3 Cumulative hazard models

The flexible parametric models which model in the log cumulative hazard scale were initially proposed by Royston and Parmar [97] and then extended for the relative survival setting by Nelson and colleagues [24].
The formulation of the model is inspired in a simple Weibull model:

$$S(t) = exp(-\xi_1 t^{\xi_2}), \tag{2.44}$$

where ξ_1 and ξ_2 are, respectively, the scale and shape parameters of the model. Using the relationship between the cumulative hazard and the survival function and taking the logarithm, this model can be written as:

$$log[\Lambda(t)] = log(\xi_1) + \xi_2 log(t). \tag{2.45}$$

The log cumulative hazard is thus a linear function of $log(t)$ in this model. Adding covariates:

$$log[\Lambda(t)] = log(\xi_1) + \xi_2 log(t) + \boldsymbol{x_i}\boldsymbol{\beta} \tag{2.46}$$

The idea of Royston and Parmar was to relax the assumption of linearity of $log(t)$ by using a smooth function to capture the baseline shape. They proposed restricted cubic splines, which have been described above. The formulation of the excess hazard model is:

$$log[\Lambda_{Ei}(t|\boldsymbol{x_i})] = s(log(t)|\gamma, \boldsymbol{k_0}) + \sum_{j=1}^{D} s(log(t)|\boldsymbol{\delta_k}, \boldsymbol{k_j})x_{ij} + \boldsymbol{x}_i\boldsymbol{\beta} \tag{2.47}$$

The $s(log(t)|\gamma, \boldsymbol{k_0})$ and $s(log(t)|\boldsymbol{\delta_k}, \boldsymbol{k_j})$ represent the restricted cubic splines functions, respectively for the baseline and for the time-dependent effects. In this formulation they are allowed to have different degrees of freedom, i.e. different number of knots and also different location.

Estimation

The contribution to the log-likelihood of the i_{th} individual is given by:

$$logL_i = \delta_i \cdot log\left[\lambda_P(t_i) + \left(\frac{1}{t_i}\frac{ds(u_i;\gamma)}{du_i} + \sum_{j=1}^{D}\frac{1}{t_i}\frac{ds(u_i|\boldsymbol{\delta_k}, \boldsymbol{k_j})}{du_i}x_{ij}\right) exp(\eta_i)\right] - exp(\eta_i), \tag{2.48}$$

where $u = log(t)$ and $\eta_i = s(log(t)|\gamma, \boldsymbol{k_0}) + \sum_{j=1}^{D} s(log(t)|\boldsymbol{\delta_k}, \boldsymbol{k_j})x_{ij} + \boldsymbol{x}_i\boldsymbol{\beta}$. Contrarily to the log-likelihood function obtained in the excess hazard models written in the non-cumulative scale, this function does not involve the numerical integration of the hazard. All functions can be obtained analytically which was pointed out by the authors as being an advantage since speeds up computation time [24].

Cumulative hazard vs non-cumulative hazard

The flexible parametric models defined in the log-cumulative scale have been proposed has having several advantages. First, under the PH assumption, the coefficients associated with the covariates can also be interpreted as log Hazard Ratios. Second, the cumulative hazard as function of log time is more stable than the hazard function being easier to capture its shape. Also, this type of scale does not require numerical integration to obtain survival or the cumulative hazard, decreasing computation time [98]. However, when there are several time-dependent effects, the interpretation of the time-dependent hazard ratios is not clear as they depend on values of other covariates, even with no interaction between these covariates, which is not the case when modelling on the hazard scale [99].

2.4.4 Model building strategies

When building an excess hazard regression model, several aspects must be taken into consideration. First, which type of model is to be used. Additive or multiplicative effect of covariates? Parametric or semi-parametric model? Cumulative or non-cumulative excess hazard scale? These different type of models were described previously. Considering the type of model chosen, then it is necessary to choose which variables must be included. The model needs to be adjusted, at least, for each life table variable to properly account for informative censoring. These are typically sex and age. The functional form of each variable (linear or non-linear) and time-dependent effects are other aspects of major importance to take into consideration. Also, interactions between variables should be investigated.

One of the criteria mostly used in model building is the Akaike Information Criterion (AIC). AIC is a statistics defined for parametric models whose parameters have been obtained by maximizing a form of likelihood function [100]. AIC values are used to compare a set of different models relatively to its fit to data. The selection is based on the minimum AIC criterion, which says that the model with smallest AIC is to be preferred. AIC is given by:

$$AIC = -2(log - likelihood) + 2K, \tag{2.49}$$

where K is the number of parameters in the model. AIC is influenced by the log-likelihood and by the number of parameters in the model. A better goodness-of-fit gives a higher likelihood and consequently a lower AIC. On the other hand, a higher number of parameters penalizes AIC. The lower AIC should thus give the model that neither under-fits nor over-fits. This criterion allows to choose the 'best' model from a set of given models but it is not a measure of goodness-of-fit neither of model quality. The Bayesian Information Criterion (BIC) is another useful statistic for model comparison. It is closely related to AIC but its penalty term (related to the number of parameters) is larger.

One alternative model building strategy is to use a iterative backward elimination procedure like the one proposed by Wynant and Abrahamowicz in 2004 [101]. First, a multivariable model including all the variables of interest and considering non-linear and time-dependent effects for all is built. Second, each of the non-linear and time-dependent effects of all the covariates are tested using likelihood ratio tests since in each pair of models being tested one model is nested in the other. At the first iteration of the process, the effect with the highest p-value is removed. The process continues, removing at each iteration the least significant of the remaining time-dependent or non-linear effects. The process ends when all the remaining effects are statistically significant. The algorithm is outlined below.

For illustration, a model with a single continuous covariate is considered. Four models can be built considering the different combinations between non-linear/linear effects of the covariate and time-dependent/proportional hazards assumption. Consider that NL stands for non-linear effects, LL for linear effects, TD for time-dependent effects, PH for proportional hazards, LRT for log-likelihood ratio test and α is the significance level chosen for the hypotheses tests.

- Fit the following models to the full dataset:
  - M1: NL + TD
  - M2: LL + TD
  - M3: NL + PH
  - M4: LL + PH
- Test for linearity and proportional hazards:
  - Test linearity assuming TD: LRT comparing M1 with M2 $\rightarrow$ p-value=p_{12}
  - Test PH assuming NL: LRT comparing M1 with M3 $\rightarrow$ p-value=p_{13}
- Eliminate least significant effect :
  - If $p_{12} > p_{13}$ and $p_{12} > \alpha \rightarrow$ eliminate NL

    - $*$ Test PH assuming linearity → choose between M2 and M4
  - If $p_{12} < p_{13}$ and $p_{13} > \alpha \rightarrow$ eliminate TD
    - $*$ Test linearity assuming PH → choose between M3 and M4
  - If $p_{12}, p_{13} < \alpha \rightarrow$ choose model M1

With more than one covariate, the process is iterative eliminating first the least significant covariate effect.

2.4.5 Goodness of fit

The goodness of fit, that is, how well a model fits the data is not easy to assess in an excess hazard regression model. Visual inspection of the model fitness cannot be achieved by directly plotting the observed vs the predicted values but it is possible to compare the predicted excess hazard or survival functions with non-parametric estimates.

Two specific points should be checked when analysing the adequacy of an excess hazard model: the proportional hazards (PH) assumption for the effects of covariates; functional form (FF) of continuous covariates. In general survival analysis when modelling hazard, the first can be done using the Schoenfeld residuals and the second using Martingale residuals [10].

Though very common in hazard models, few methods are available for testing the assumption of proportional hazard assumption and the functional form of covariates in excess hazard models. Stare and colleagues [102] proposed a test based on Schoenfeld-like residuals and on Brownian bridges to test the PH assumption. The test is available in the R statistical package $relsurv$. A few years later, Cortese and colleagues proposed a new approach for goodness of fit of excess hazard models, which consisted on statistical and graphical tests based on cumulative martingale residuals [103]. Recently, Danielli and colleagues proposed new formal tests to check the proportional hazard assumption and the functional form of covariates also based in cumulative martingale residuals [104]. These tests were being implemented in the $mexhaz$ package from R, but were not available yet.

2.5 Age-standardized net survival

Net survival generally depends on age. The comparison of net survival estimates from two or more regions or periods can be misleading if the age distributions of the populations to be compared are very different. It is thus necessary to control for the differences in the age structure of the populations in comparison, age being implicitly considered as the only factor influencing net survival.

If two populations observed in two different regions (say A and B) are exposed to the same age-specific survival probabilities, the age-standardised survival obtained from each region, $S_A^w(t)$ and $S_B^w(t)$ should be equal:

$$S_A^w(t) = \int w(age) S_A(t|age) dH(age)$$

$$S_B^w(t) = \int w(age) S_B(t|age) dH(age), \tag{2.50}$$

where H is the distribution function of age. An age-standardised estimate can be interpreted as the survival one population would have if its age structure was the same as the age structure of a standard population. In order to compare different age-standardised estimates, the standard population used must be the same for all populations in comparison. The age-standardised survival does not reflect the actual survival of any population. It is a measure just useful for comparisons.

The age-standardisation of net survival is usually performed using a discrete version of the equations presented above (2.50). Using a discrete age distribution, the age-standardised net survival ($ASNS$) is given by a weighted mean of age group specific survival (2.51).

$$ASNS(t) = \sum_{j=1}^{k} w_j \cdot S_{Nj}(t), \tag{2.51}$$

where $S_{Nj}(t)$ is the net survival estimate of age group j and w_j ($\sum_{j=1}^{k} w_j = 1$) is the corresponding weight. The standardisation thus requires the estimation of net survival for each age group.

Three questions arise from equation (2.51). Which weights to use, what age categorization and how to estimate net survival for each age class. On one hand, a thinner age categorization would be a better approximation of the integral. On the other hand, narrower age groups can cause stability issues in the age group-specific net survival estimates.

In population-based cancer survival analysis an International Cancer Survival Standard, proposed in 2004 [1], is used by most of the cancer registries and in international comparison studies. Five age groups are considered in these standard populations and the set of weights depend on the type of cancer (Table 2.1). The standards were derived from the European survival study EUROCARE-2 dataset and aimed at minimising the difference between the raw and the standardised estimates.

To estimate $ASNS(t)$ from age group specific net survival estimated non-parametrically, it is necessary to use equation (2.51). The $S_{Nj}(t)$ are calculated using the PP estimator. The variance of the age-standardised net survival is given by:

$$VAR(ASNS(t)) = \sum_{j=1}^{k} w_j^2 \cdot VAR(S_{Nj}(t)), \tag{2.52}$$

assuming independence between the age group-specific survival and $VAR(S_{Nj}(t))$ being the variance of the net survival for age group j.

If net survival is estimated from a multivariable excess hazard model, the age-group specific net survivals are obtained by averaging the individual net survival predictions ($S_i(t)$):

$$S_{Nj}(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} S_i(t) \tag{2.53}$$

where n_j is the number of patients in age-group j. The individual net survival are obtained by integrating the excess hazard function:

$$S_i(t) = exp\left\{ - \int_0^t \lambda_{E_i}(u)du \right\} \tag{2.54}$$

Age group	ICSS 1 [1]	ICSS 2 [2]	ICSS 3 [3]
15-44	0.07	0.28	0.60
45-54	0.12	0.17	0.10
55-64	0.23	0.21	0.10
65-74	0.29	0.20	0.10
75+	0.29	0.14	0.10

Table 2.1: International Cancer Survival Standards for broad age groups [1].

The derivation of the variance for age-standardised net survival based on model predictions was made in the scope of the study presented in Section 3.1. In this Section the question of age-standardisation is addressed, with special emphasis to situations where data are sparse.

[1] Lip, tongue, salivary glands, oral cavity, oropharynx, hypopharynx, head and neck, oesophagus, stomach, small intestine, colon, rectum, liver, biliary tract, pancreas, nasal cavities, larynx, lung, pleura, breast, corpus uteri, ovary, vagina and vulva, penis, bladder, kidney, choroid melanoma, non-Hodgkin lymphomas, multiple myeloma, chronic lymphatic leukaemia, acute myeloid leukaemia, chronic myeloid leukaemia, leukaemia, prostate

[2] Nasopharynx, soft tissues, melanoma, cervix uteri, brain, thyroid gland, bone

[3] Testis, Hodgkins disease, acute lymphatic leukaemia

2.6 Socioeconomic inequalities in cancer survival

Cancer survival has been shown to vary most of the times with socio-economic group. Kogevinas and Porta in 1997 [35] and, in 2006, Woods and colleagues [3], reviewed various studies for different tumour sites and populations (England, Scotland, Canada, United States, Australia, Norway, ...), showing evidence for survival differences between socio-economic groups. Specifically for colorectal cancer, two review studies done in 2010 and 2014 [4, 5] present a considerable amount of literature favouring the evidence of worse survival for patients with a lower Socioeconomic status (SES). Other studies, however, did not find evidence of SES inequalities in survival from cancer [105, 106, 107]. Methodologically, several differences can be found among the literature evaluating SES inequalities in cancer survival. The measure used as outcome can differ (overall survival, relative survival, hazard ratio, ...) as well as the indicator of SES. Simple or composite indices, attributed at individual or area level are options differing from study to study.

2.6.1 Socioeconomic indices

SES indicators can be simple, if based in a single measure, or composite if they are the result of the combination of different single indicators. Table 2.2 presents some of the most used single indicators with a brief description and some examples of references of studies evaluating cancer survival inequalities where they have been used. Some of the presented indicators are measured at individual level while others are area-based.

The categories used, not only the number but also the cut-offs defined, in each of the categorical indicators as, for instance, education vary between studies. This heterogeneity hampers the comparability between different studies. Single indicators are easy to calculate but have the disadvantage of reflecting only certain aspects of deprivation. Composite indices try to condense in a single measure the several dimensions of deprivation. Carstairs and Townsend scores are two of the most classical composite indices [39, 38]. The variables involved in each one are presented in Table 2.3. In both indices, all variables have the same weight for the final score. Gordon [117] argues that indices that attribute equal weight to their component variables 'are likely to yield inac-

Indicator	**Description**	**Cancer survival studies**
Education	Level of education attained (categorical) or total number of years of education	Egeberg *et al.* (2008) [108] Hussain *et al.* (2008) [109] Dejardin *et al.* (2006) [110]
Income	Usually household income, adjusted or not for household size. Categories depend on country	Dejardin *et al.* (2006) [110] Kim *et al.* (2011) [111] Gorey *et al.* (2011) [112]
Unemployment	Lack of employment	Ueda *et al.* (2006) [113] Dalton *et al.* (2008) [114]
Occupation	Different occupational-based indicators exist. Reflect the type of work the patient has	Egeberg *et al.* (2008) [108] Auvinen *et al.* (1995) [115] Kravdal Ø(2000) [116]

Table 2.2: Single indicators of SES most used in cancer survival research (based on [2, 3, 4, 5]).

curate results'. The author suggests that the components should be weighted to reflect the different probability that each group has of suffering from deprivation. The Index of Multiple Deprivation is nowadays most commonly used in the UK for measuring levels of deprivation. The overall index is a combination of seven indices each measuring different domains of deprivation. The indices are based on routine administrative data (and not on census data) and are regularly updated.

The European Deprivation Index (EDI) was first proposed by Pornet and colleagues in 2012 [42]. The index is based on census variables available for each country that are most associated with variables identified from the European Union Statistics on Income and Living Conditions (EU-SILC) survey [118]. The index was first developed for France and then applied to other European countries, namely, England, Italy, Portugal and Spain [40]. Later, it has been developed also for Slovenia [41]. The index for Portugal was based on 2001 census and includes percentage of: non-owned households, households without indoor flushing, residents with low education level (6th grade), household with 5 rooms or less, unemployed looking for a job, female residents aged 65 years or more, households without bath/shower and percentage of residents employed in manual occupations [46]. A continuous score was obtained for each census tract based on the census responses of its inhabitants. The index is also available at parish and municipality level. This score can be categorised in deciles or quintiles.

Index	Description	Cancer survival studies
Carstairs	Based on four census variables: non-car ownership, overcrowding, unemployment and social class.	Coleman et al. (2004) [119] Mitry *et al.* (2008) [120]
Townsend	Based on four census variables: non-car ownership, overcrowding, unemployment and non-home ownership.	Pollock *et al.* (1997) [121] Lejeune *et al.* (2010) [122]
Index of Multiple Deprivation (IMD)	Combination of seven distinct dimensions of deprivation called Domain Indices: Income; Employment; Health and disability; Education, skills and training; Barriers to Housing and Services; Living environment; Crime.	Fowler *et al.* (2017) [123] Abdel-Rahman *et al.* (2014) [124] Shafique & Morrison (2013) [125]
European Deprivation Index (EDI)	Weighted combination of aggregated variables from the national census that are most highly correlated with a country-specific individual deprivation indicator.	Di Salvo *et al.* (2017) [126] Belot *et al.* (2018) [127]

Table 2.3: Composite indices of SES most used in cancer survival research (based on [2, 3, 4, 5]).

The single indicators presented (education, income, ...) can be measured at individual level [108] or based on the area of residence at diagnosis of the cancer patient [110]. The composite indices presented are built at area level and reflect the condition of geographical areas.

The level of deprivation that each individual is subjected to is influenced by two factors: individual and ecological. Sloggett and Joshi [128], when studying the association between socioeconomic level and health indicators, have shown that even after adjusting for individual level, the ecological effect is still significant, at least for some indicators. Diez Roux [129] discusses the importance of considering group-level variables besides individual level ones since both can reflect different types of health conditioners. However, in population-based studies, individual information on SES is not commonly available, at least in Portugal. So, in the absence of individual information, the socioeconomic condition must be attributed according to the condition of the patients geographical area of provenance. The use of area based indicators or indices as proxies for individual patient condition when individual measures are not available should be done with caution. To better reflect the individual condition, the geographical areas should be preferably small and homogeneous regarding socioeconomic conditions. The estimates of the association between SES and cancer survival, or in general the health outcome, can be underestimated relatively to the true individual-level effect since all individuals in a certain area are given the same score diluting possible differences [2].

The use of different SES indicators or indices can lead to find different associations between deprivation and the outcome of interest. However, Woods and colleagues [37] observed, for breast cancer, that the deprivation gap in survival was more influenced by the population size of the geographic area used for the attribution of the socioeconomic indices than by the definition of the index.

2.6.2 Attribution of SES to cancer patients

The attribution of a SES condition to a patient using area-based indices can be done using different geographic units. In England and Wales the smallest units are the *Lower-layer Super Output Areas*, with a mean population of 1500 inhabitants [130]. Each unit has a corresponding postcode making it possible to match the patient area of residence with the one for which the index is defined using that information. In France, the territory is divided in areas with a target size of 2000 residents per basic unit designated *IRIS* (acronym in French for 'aggregated units for statistical information') [131].

In Portugal, the smallest geographical unit for which there is statistical information is called *sub-secção estatística*. In urban areas corresponds to a city block. The second smallest unit is termed *secção estatística*. Corresponds to a census tract with about 300 accommodations [132]. The Portuguese version of the EDI is available for this last geographic unit. There is no direct correspondence between postcodes and census tracts. To match the patient's address with the corresponding geographic unit for which the SES index is available it is necessary the aid of geographical information software. First, each address must be geocoded, i.e. XY coordinates must be associated to the address. Then, the addresses mapping must be overlaid with the deprivation distribution to make the correspondence between both.

2.6.3 Possible reasons for SES inequalities

Several different reasons can contribute to explain socioeconomic inequalities in survival from cancer. Health awareness and screening participation can be higher in more affluent groups leading to an earlier diagnosis of the disease. This advance in diagnosis can have a real impact on prognosis or not. Earlier diagnosis can just (artificially) increase survival time by advancing the time of diagnosis but without delaying time of death (lead-time bias) [115]. In other more favourable situations, an earlier diagnosis can mean, in fact, that the disease is diagnosed in an earlier stage making possible different treatment intention that will favour prognosis.

Differential access to treatment between less and more deprived groups can also justify inequalities in survival [4]. The type of treatment applied may also be a explanatory factor. However the type of treatment can be related with other factors such as stage at diagnosis or comorbidities that can have themselves heterogeneous distributions.

Different tumour characteristics, possibly caused by different aetiology factors, can also lead to inequalities in the outcomes. Last, host factors can also help explain inequalities, namely, different comorbidities or health behaviours before or after the disease has been diagnosed [115].

2.6.4 Survival measures

The outcome used in studies evaluating the association between SES and survival from cancer is not homogeneous. Some studies used overall survival as measure [112, 133, 111]. Overall survival is not only influenced by the disease in study but also by the other causes mortality. SES inequalities can be wrongly attributed to cancer when using this type of survival measure. Cancer-specific survival is another measure that is found on the literature [134, 135]. This type of measure requires good quality information on cause of death, which is seldom available in population-based studies. Relative survival has been used also [108, 136, 120]. In this setting, the survival attributable to the disease is based on the comparison of the observed survival with the background mortality of a matched population. Since SES can also affect mortality from other causes, the background mortality should be adjusted for these factors. Otherwise, overestimation of the effects of SES in cancer survival inequalities can occur [115].

2.7 Life tables

Life tables provide information on mortality rates and probabilities of death of specific populations defined by geographical regions or periods of time. They are an important demographic tool as they are the basis for the estimation of life expectancy at birth, an important population health status indicator. Life tables are usually stratified by age, sex and calendar year. Other factors have been shown to influence also population general mortality such as deprivation status or ethnicity.

Life tables are fundamental in the estimation of survival in the relative survival setting [66]. They provide information on background mortality of cancer patients necessary to estimate the survival attributable to the disease in study. To allow the unbiased estimation of cancer related survival, these tables should correctly represent the population mortality from which the patients were drawn from.

Life tables are built using counts of deaths and population at risk stratified by relevant demographic variables (age, sex, others). This information is usually made available by the national statistics offices. It is also possible to find some web sites that aggregate mortality information from several countries. The Global Health Observatory data repository from WHO [137] and the Human Mortality Database [138] are two examples.

Life tables rely on the analysis of a fictional generation submitted to the mortality rates observed during a certain specified period. This generation is assumed to be a closed cohort where drop-outs are only possible by death (no migrations are allowed). Several variables are typically represented in these tables (2.4).

age	m_x	q_x	l_x	d_x	L_x	T_x	e_x
0	0.00543	0.00541	100000	541.2	99729.4	8061894.2	80.6
1	0.00054	0.00054	99458.8	53.8	99431.9	7962164.9	80.1
2	0.00037	0.00037	99405.0	36.7	99386.7	7862733	79.1
3	0.00037	0.00037	99368.3	37.1	99349.7	7763346.3	78.1
...	...	...	...	...	...	...	...

Table 2.4: Example of a life table.

The quantities in the table are the age-specific mortality rate (m_x), the probability of dying in an age interval $x \rightarrow x + \Delta x$ (q_x), the number of persons in age class x of the fictional population (l_x), the number of deaths in age class x (d_x), the number of person years lived between age x and $x + \Delta x$ (L_x), the total number of person years lived after age x (T_x) and the life expectancy at age x (e_x).

In complete life tables Δx represents single years of age while in abridged life tables it is usually 5 years. Besides age, life tables are usually stratified by sex and calendar year but can also be stratified by any variable that affects background mortality. It is possible to find for regions, such as US or UK, life tables stratified by socioeconomic condition and ethnicity.

Information necessary to build life tables include number of births, number of deaths by age and population distribution by age. Each of these variables stratified by relevant variables (eg. sex, calendar year, region, $\cdots$).

Considering that the mortality rate by age class is known, the other quantities in the table can be calculated as follows.

Assuming that the instantaneous mortality rate remains constant throughout the age interval from x to $x + \Delta x$, the probability of death can be derived from the mortality rate using:

$$q(x) = 1 - exp[-m_x] \tag{2.55}$$

The number of survivors at age x (l_x) is given by:

$$l_x = l_{x-1}(1 - q_{x-1}) \tag{2.56}$$

Life tables are built assuming a closed population (i.e. no migrations) and with fictitious starting population (l_0) of 100,000. The distribution of the number of deaths by age (d_x) is given by:

$$d_x = l_x \cdot q_x \tag{2.57}$$

The total number of years (L_x) lived by survivors l_x on the interval $[x, x + \Delta x[$ is given by:

$$L_x = l_x - \frac{1}{2}d_x, \tag{2.58}$$

assuming that deaths occur uniformly on the age interval $[x, x + \Delta x[$. The total number of years lived by the population after age x, can be obtained by:

$$T_x = \sum_{t=0}^{w-1} L_{x+t}, \tag{2.59}$$

where w represents the maximum age attainable on the life table. Life expectancy at age x is calculated by the ratio of this quantity and the number of survivors:

$$e_x = \frac{T_x}{l_x} \tag{2.60}$$

Life expectancy at birth is obtained making $x = 0$.

To first calculate death rates, it is necessary to have information on deaths and population at risk stratified by the relevant variables. This information is not always available in single year intervals neither is available until the last age of the life table (100 years or more). Interpolation, or extrapolation for the oldest ages, needs to be used to obtain mortality rates by intervals of one year. Several methods for building complete life tables from abridged data have been in use, namely, Elandt-Johnson, Kostaki, Brass logit, and Akima spline methods [139]. More recently, Rachet and colleagues [140] suggested a modelling approach to estimate smoothed mortality rates using flexible Poisson multivariable models. Death counts are modelled in the generalised linear model framework, considering a Poisson error and using splines to capture the effect of age. Considering just the effect of age, the model is given by

$$log(d_x) = \beta_0 + f(x) + log(pyrs_x), \tag{2.61}$$

where d_x represents the number of deaths for age x and $pyrs_x$ the person-years at risk for the same age. Men and women are typically modelled separately since the mortality rates of both sexes are significantly different. This method can use complete or abridged raw data allowing the estimation of complete life tables. This type of models was considered recommendable because it derives robust and unbiased estimates without making strong assumptions about age-specific mortality profiles. Also, a simulation study has shown that this method had better goodness-of-fit performance than other implemented methods [140, 141].

In Portugal, no deprivation-specific life tables are available precluding the correct estimation of deprivation-specific survival. This question was addressed in Section 3.3.

2.8 Missing data

2.8.1 The issue with missing data

In epidemiological studies a set of collected data usually consists in a number of rows representing subjects or cases and a number of columns each corresponding to a different variable measured for each case. If this matrix is not totally filled then it means that there are missing data. That is to say that for certain subjects, some values of one or more variables were not recorded for some reason.

Missing information is a very common issue in observational studies. This missingness can have multiple causes. People may not answer all questions in a questionnaire, a registrar can forget to record some information, some periodic evaluations can be missed in longitudinal studies due to patient absence, records can be lost, clinical characteristics can not be measured due to patient condition, besides many other examples. In these situations the values are not observed/recorded but could have been if there were no failures in the information collection process. Missing data can thus be defined as data that actually exists but was not observed. A broader definition of missing data can be thought including non-observable potential outcomes. For example, what would be the outcome of a patient in a clinical trial had they been chosen to a different arm of treatment. The present study deals with missing data in the sense of the first definition.

The easiest way of dealing with this issue is to ignore all the cases that have at least one variable with missing information. In datasets with large number of different variables, this can lead to a substantial reduction of the sample size available for analysis and a large waste of information. The obvious consequence is a reduction of statistical power, that is, the ability to find differences when they exist. Also, depending on the reasons why the data are missing, performing the analysis using only the fully observed cases may lead to biased results.

2.8.2 Missing data in cancer survival analysis

In the context of survival analysis, missing data can occur both in the outcome of interest (survival time, status) as well as on the variables that help explain survival (covariates). These can include morphology, grade, stage, biomarkers status, comorbidities, socioeconomic status among many others [27, 142]. In population-based cancer survival analysis, the existence of missing information in key prognostic factors is also a general issue [143, 144, 145]. Stage of disease at diagnosis, one of the most important cancer prognostic factor, is a variable that has usually a considerable proportion of missing information. Stage of disease can be missing from the cancer registries because it has not been actually assessed or because it was assessed but not properly recorded. Several studies that analysed the causes of stage missingness can be found in the literature [144, 146, 147, 148]. The proportion of cases with unknown stage can vary considerably between cancer registries but also between different tumour sites. Gurney and colleagues [146] analysed factors that affected the chances of stage being missing from the records of the New Zealand Cancer Registry for 18 different tumour sites. A range from almost no missing cases for testis cancer up to 73% for prostate cancer was observed. A similar study in the United States, found much lower stage missingness percentages, with a maximum below 30%. The tumour sites with the highest proportion were bad prognosis cancers such as liver, pancreas and oesophagus [149]. Due to clinical reasons, staging can be more or less difficult to evaluate and also more or less useful depending on the tumour site. Tumour sites for which surgery is the primary treatment can be less likely to have missing stage. It has been observed that patients with higher level of comorbidities and poor health status have higher chance to have unknown stage although its effect depends on cancer site [146, 149]. Socioeconomic condition, area of residence,

ethnicity and type of Health Service where diagnosis is performed were factors that have also been found to influence the likelihood of stage recording. Completeness of tumour staging information seems to generally decline with age [144, 146, 147, 148]. Worthington and colleagues [147] found that in the US African Americans colorectal patients had less chances of having known stage. In New Zealand the same was observed for Maori patients diagnosed with lung and cervical cancers but not for other cancers [146].
The reasons for having missing information are therefore multifactorial. The impact of the missing data on the results of statistical analysis depends on the mechanism which caused the data to be missing. The terminology used in the classification of the missing data mechanisms is presented below.

2.8.3 Missing data patterns

Missing data patterns concerns to the way missingness occurs in the variables of a certain dataset. It has implications in the methods that are employed for handling that missingness. The pattern of missing data is designated by *monotone* if it is possible to re-order the, say p, variables in a matrix such that for every line (case) i and column (variable) j:

- for case i, the variable j was observed ($j = 2, \cdots, p$), and for this case all variables $j' < j$ where also observed, and
- for case i, the variable j is missing ($j = 2, \cdots, p$), and for this cases all variables $j' > j$ are also missing.

This often occurs in longitudinal studies, when there is drop-out. In cancer survival analysis settings, it is more natural to find non-monotone missing data patterns.

2.8.4 Missing data mechanisms

The mechanisms that lead to data being missing concerns with the relationship between missingness and the values of variables [28]. It has direct implication on the way missing data should be handled. The concept of missing data mechanism was first formalized by Rubin in 1976 [150]. Rubin considered the missing-data indicators as random variables and assigned them a distribution. Three different missing data mechanisms are nowadays commonly accepted in the literature: Missing completely at random (MCAR), Missing at random (MAR) and Missing not at random (MNAR) [28]. To formally distinguish between these three mechanisms, let us first introduce some notation. Let $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,p})^T$ denote the set of p variables that are intended to be collect for case i. Let $\mathbf{Y}_{i,obs}$ denote the subset of the p variables that was observed for each case ($i = 1, \cdots, n$). Let $\mathbf{Y}_{i,mis}$ denote the subset of the p variables that are missing. The set of observed and missing variables can thus be different from case to case. Let $\mathbf{R}_i = (R_{i,1}, \cdots, R_{i,p})^T$ denote the missing indicator such that $R_{i,j} = 1$ if $Y_{i,j}$ is observed and $R_{i,j} = 0$ if $Y_{i,j}$ is missing [151]. The missing data mechanism, defined as the conditional probability $P(\mathbf{R}_i|\mathbf{Y}_i)$, is classified as:

- *Missing Completely At Random (MCAR)*
  If the probability of having a missing value is not dependent on the observed or on the missing values, i.e.
  $$P(\mathbf{R}_i|\mathbf{Y}_i) = P(\mathbf{R}_i)$$
- *Missing At Random (MAR)*
  If conditional on the observed data, the probability distribution of $\mathbf{R}_i$ is independent of the unobserved data, i.e,
  $$P(\mathbf{R}_i|\mathbf{Y}_i) = P(\mathbf{R}_i|\mathbf{Y}_{i,obs})$$
- *Missing Not At Random (MNAR)*
  If the missing data mechanism is neither MCAR nor MAR, then it is classified as MNAR. This means that, even conditional on the observed values, the probability of a value being missing depends on the unobserved value itself:
  $$P(\mathbf{R}_i|\mathbf{Y}_i) \neq P(\mathbf{R}_i|\mathbf{Y}_{i,obs}).$$

It is not possible from the observed data to infer which type of mechanism caused a set of data to have missing data. When dealing with missing data methods it is necessary to make assumptions regarding the missingness mechanism that are not testable. Nevertheless, the available data can be used to help in the formulation of plausible assumptions and on the choice of methods to handle the missing data.

2.8.5 Methods for dealing with missing data

A brief overview of several common approaches to deal with missing data are presented next.

Avoid missing data

The obvious way of not having to deal with missing data is not having them. Unfortunately, in most of the electronic health records data sets it is very difficult to fully observe all variables. In the context of population-based cancer survival analysis, this is even harder. Population-based cancer registries collect information from a wide set of different sources, namely, public and private hospitals and pathology laboratories, making information recovery a difficult task.

Complete-case analysis

Complete-case analysis consists in limiting the analysis to the cases for which all variables were observed. This is the most simple approach to deal with the occurrence of missing data and is the one most statistical packages adopt by default (*listwise deletion*). In situations of non-monotone missing data pattern, with missing information in several variables, this can result in a substantial sample size decrease. This loss of information has two consequences. First a decrease in statistical power due to the reduction of the number of cases available for analysis. Second, if the missing data mechanism is not MCAR, bias in the results of the analysis [152]. For some specific situations, under MAR or MNAR the complete-case analysis can be unbiased [153], although the practical application of that result is limited since in real world situations the true missingness mechanism is not know.

Complete-case analysis can be a reasonable approach if the proportion of incomplete information is small (some authors say 5%), although there are no rules of thumb on the maximum acceptable proportion of missing cases since the impact on the analysis is dependent on several factors [28].

Inverse Probability Weighting

In the Inverse probability weighting (IPW) approach, the analysis model is fitted only to complete cases, but different weights are given to each case. The weight is inversely proportional to the probability of a case being observed. That is to say that cases with higher probability of being missing have a higher weight in the model in order to correct for the bias that the complete-case analysis would introduce. The probabilities of missingness are obtained from the data, not only from the outcome and the set of covariates that one includes in the analysis model but also from any further variables available. Although IPW performs worse than other methods, Seaman and White [154] argued that it can be a valuable approach in certain settings as long as care is taken to ensure that the missingness model is correctly specified and that weights are not unstable.

Indicator method

In this approach, no subjects are excluded from the analysis. For each variable that is not fully observed, an extra indicator variable is created. This new variable takes the value 1 for the cases in which the original variable is missing and 0 otherwise. Let X be the original variable and M the missing indicator variable. Then in the analysis model, the original variable should be replaced by $(1 - M_i) \cdot X_i$ and the extra M_i. For categorical covariates this is equivalent to create an additional ’missing’ category for that variable. These categories can group a set of very heterogeneous classes into a single group. This method can produce severe biased results and has been criticised and discouraged [155, 156]. In some situations, however, such as missing data in randomised trials baseline covariates, the method can produce unbiased estimates [157]

Single imputation

A simple approach for handling missing data is the single imputation method. It consists in filling each missing value with a likely value. Different single imputation techniques exist. One of the most simple and common technique is the mean imputation, for continuous variables with missing values, or the mode imputation for categorical variables. The mean can be estimated conditional on other variables or it can be simply the mean (unconditional) of the variable with missing values. This method has the advantage of being easy to implement, produces stable imputed values and does not require any distributional assumption [158]. It requires the assumption that the missing data mechanism is MCAR. After imputing the missing data, the completed dataset is treated in the analysis step as if all data has been observed. The uncertainty related to the imputed values is not taken into account. This results in an underestimation of the variance of the analysis model coefficients and consequently too narrow confidence intervals or too small p-values. On the other side, imputing the overall mean in covariates will dilute its association with the outcome and bias regression coefficients towards the null [159].

Another single imputation technique is hot-deck imputation. In this approach, the imputed value is randomly selected from the set of fully observed values that share the same covariate values [28]. Relatively to the mean imputation, it introduces more variability since two missing values in a certain variable that have the same values of all other variables can be imputed with different values. It has also the advantage of imputing only possible values for each variable as the imputed value is selected from the set of observed values. It does not make assumptions about the distributional form of the missing values although it is necessary to select which variables are used to match cases.

In all the single imputation techniques presented, the uncertainty of the imputations is not taken into account. The imputed values are indistinguishable from the observed ones. To take this uncertainty in consideration, several imputations are needed for each missing value. This is what is done in the multiple imputation approach presented next.

Multiple imputation

The approaches presented above are simple to apply but will generally result in misleading conclusions. Multiple imputation has become in the last years a popular approach for accommodating incomplete information in statistical analysis. This method is not exempt of its limitations and inherent assumptions. When misused, it can also lead to biased conclusions. In the present study, attention has been focused on this approach for dealing with missing data. The method is discussed in more detail below.

2.8.6 Multiple imputation methods

Multiple imputation (MI) was first introduced by Rubin in 1978 [160]. Initially, MI was developed in the framework of survey nonresponse but has nowadays been expanded to a broader set of different fields, including survival analysis [151].

In MI several imputations are generated for each missing value, as opposed to single imputation where each missing value is replaced by a single value. This creates several completed datasets, as many as the number of imputations performed. Each completed dataset is analysed using standard methods for complete data. The results from the several analyses are then combined to produce single estimates and confidence intervals that incorporate missing-data uncertainty. The objective in MI is not to estimate the values that are missing but to obtain valid inference from the completed datasets.

The process can be divided in three main steps: the imputation, the analysis and the combination steps. The models related to the first step are commonly designated as imputation models and the ones used in the second step, as substantive (or analysis) models [29]. One of the advantages of MI is that the imputation model can include more variables than the substantive model. Independently of the imputation and substantive models used, briefly the algorithm goes like this (Figure 2.2):

i. Using the imputation model, generate $M > 1$ values for each missing value, obtaining M completed datasets;

ii. Fit the substantive model independently to each one of the M completed datasets;

iii. Combine the results obtained from each analysis performed in the previous step using Rubin's rules (defined below).

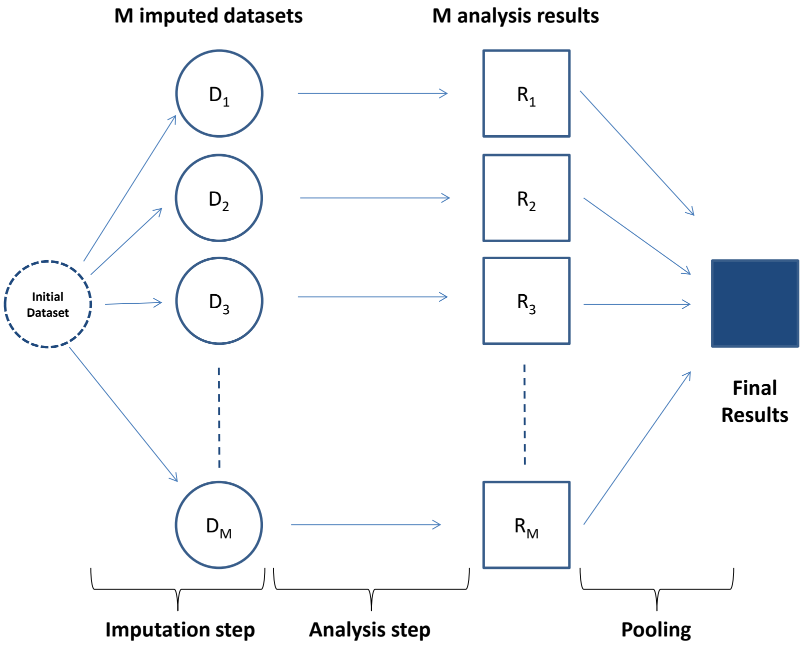

Figure 2.2: Schematic representation of the MI method (adapted from [7])

Multiple imputation uses Bayesian inference to sample the missing values from their posterior predictive distribution. Assuming that $Y = (Y_{obs}, Y_{mis})$ has a parametric model $P(Y|\theta)$, that θ has a priori distribution $P(\theta)$ and that the missing mechanism can be considered MAR, i.e. the probability of missingness does not depend on unobserved information, then since by the Bayes theorem:

$$\underbrace{P(\theta|Y)}_{\text{posterior}} \alpha \underbrace{P(Y|\theta)}_{\text{model}} \times \underbrace{P(\theta)}_{\text{prior}},$$

to obtain imputations for the missing values:

i. First, draw a sample of the unknown parameters from their observed-data posterior distribution:

$$\theta^* \sim P(\theta|Y_{obs})$$

ii. Second, given θ^*, draw a random sample of the missing observation from the conditional predictive distribution:

$$Y^*_{mis} \sim P(Y_{mis}|Y_{obs}, \theta^*)$$

To draw samples from the observed-data posterior which is not typically a standard distribution, Markov chain Monte Carlo (MCMC) methods can be used. MCMC methods comprise a class of algorithms for sampling from a probability distribution.

MCMC is used to generate pseudorandom draws from multidimensional probability distributions via Markov chains that would be otherwise intractable. A Markov chain is a sequence of random variables in which the distribution of each element depends on the value of the previous one. By constructing a Markov chain that has the desired distribution as its equilibrium distribution, a sample of the desired distribution can be obtained by observing the chain after a number of steps. The process of building the Markov chain is iterative. In the first step, the missing values are sampled from the conditional predictive distribution $Y^{(t)}_{mis} \sim P(Y_{mis}|Y_{obs}, \theta^{(t-1)})$. Then, in the second step, the unknown parameters are sampled from a simulated complete-data posterior $\theta^{(t)} \sim P(\theta|Y_{obs}, Y^{(t)}_{mis})$. The first step is repeated using the new estimates of θ. The iterative process starts with given initial values ($\theta^{(0)}$), creating a Markov chain $(Y^{(1)}_{mis}, \theta^{(1)})$, $(Y^{(2)}_{mis}, \theta^{(2)})$, ... This should converge in distribution to $P(Y_{mis}, \theta|Y_{obs})$. After a certain number of initial iterations, necessary to stabilise the estimates, random samples of the missing values can be drawn from the built chain [161].

After obtaining the M imputed datasets, each is analysed using a substantive model and M different estimates of the parameter of interest ($\hat{\beta}_j$) and its corresponding variance (s^2_j) are obtained. To combined these results into an overall MI estimate and standard errors to provide valid statistical results, Rubin [30] developed a set of rules, now commonly designated as *Rubin's rules*. The MI estimator of the parameter of interest is given by averaging the individual estimators:

$$\overline{\beta} = \frac{1}{M}\sum_{j=1}^{M}\hat{\beta}_j$$

The estimated variance combines two sources of variation. The within variability, obtained by averaging the individual variance estimates:

$$\overline{U} = \frac{1}{M}\sum_{j=1}^{M} \hat{s}_j^2,$$

and the between-imputation variability that represents the variation across imputations and is related to the uncertainty caused by the missing values:

$$B = \frac{1}{M-1}\sum_{j=1}^{M} \left(\hat{\beta}_j - \overline{\beta}\right)^2.$$

The total variance is given by the sum of both variances. The multiplier $(1 + M^{-1})$ is a bias adjustment for small M:

$$T = \overline{U} + (1 + M^{-1})B$$

This estimator standardised follows, approximately, a t_ν distribution:

$$\frac{\hat{\beta} - \beta}{\sqrt{T}} \approx t_\nu$$

where the degrees of freedom (ν) are given by:

$$\nu = (M-1)\left(1 + \frac{\overline{U}}{B}\right)^2.$$

Thus, inference on the parameters can be made using this distribution.

Multiple imputation approaches can be divided into two general frameworks: joint model (JM) imputation and fully conditional specification (FCS) imputation, also known as Multiple Imputation Chained Equations (MICE) or sequential regression multivariate imputation. Both frameworks assume that a multivariate joint distribution for the data exists. While JM draws missing values for all incomplete variables in a single step from that joint distribution, the FCS approach imputes each variable at a time, drawing missing observations from a series of univariate distributions without specifying the joint distribution [162].

Using the JM strategy, a joint distribution for all variables of interest in the data must be specified. Imputations for variables with missing data are drawn from the corresponding conditional distributions, given all other variables. The method can be difficult to implement when the number of variables is large and different types of variables exist (continuous, categorical). Most software that have this strategy implemented, assume that the data follows a multivariate normal distribution [163].

To overcome some limitations and difficulties of the JM approach, an alternative method to perform MI was introduced in the late 90's and is now very popular [162]. The fully conditional specification approach splits a k-dimensional problem into k one-dimensional problems. For each variable with missing values, a distribution conditional on all other variables is specified ($P(Y_j|Y_{-j},\theta_j)$), where $Y_{-j}=\{Y_1,\cdots,Y_{j-1},Y_{j+1},\cdots,Y_k\}$. This has the advantage of being possible to specify different models for each variable, offering a greater flexibility in the choice of models. The main issue with FCS is that the implied joint distributions may not exist theoretically and convergence criteria are ambiguous [164].

One iteration (say the t^{th}) of the FCS approach for multivariate missing data consists on the following successive steps [164]. For the first variable with missing values, a draw of the parameters (θ_1) and the missing value is made from the conditional distribution $P(Y_1|Y_{-1})$:

$$\theta_1^{*(t)} \sim P(\theta_1|y_1^{obs}, y_2^{t-1}, \cdots, y_k^{t-1}),$$

$$y_1^{*}(t) \sim P(y_1^{mis}|y_1^{obs}, y_2^{t-1}, \cdots, y_k^{t-1}, \theta_1^{*(t)}).$$

To impute values for the variable y_1 in this iteration, the values imputed to the other variables in the previous iteration are used. The same procedure is executed to all variables with missing values, what constitutes one iteration. The process follows, starting again in the first variable and repeating the procedure for all. A number of iterations are run for the algorithm to 'converge'. This gives one set of imputed values. All the process must then be repeated to obtain further imputations.

The FCS approach only corresponds to imputation from a well defined joint model in some special situations. It is possible that the univariate specified models are incompatible with each other.

Some studies can be found in the literature comparing both methods. Van Buuren [165] concluded that FCS is a useful and easily applied flexible alternative to JM when no convenient and realistic joint distribution can be specified. Lee and Carlin [166] found similar behaviour between FCS and JM using multivariate normal distributions, even in the presence of binary and ordinal variables.

Number of imputations needed

Rubin [30] argued that a small number of imputations (2 to 5) is needed to obtain efficient estimates of the parameters of interest. Rubin shows that the relative variance of using only M imputations instead of an infinite number is approximately $(1 + \lambda/M)$, where λ is the fraction of missing information. However, this relates only to the point estimates of the parameters and not to the precision of the standard error estimates. Carpenter and Kenward [151] advocate to do at least $M = 100$ imputations to obtain acceptable errors. Nowadays, as computers evolved and allow much faster calculations than what was possible when these methods were first proposed, a large number of imputations is preferred.

Imputation models

Multiple imputation provides valid estimates under the MAR assumption and provided that imputations are drawn from an appropriate distribution. To increase the chance that the missingness depends only on observed data, the maximum possible number of predictors of missingness should be included even if they are not to be included in the final analysis. If a variable is predictive of missingness in another variable that is being imputed, that variable should also be included in the imputation model. Even if a variable is not predictive of missingness, but it is predictive of the partially observed variables, it should be included in the imputation model in order to reduce the uncertainty in imputing missing values, thus increasing statistical efficiency [167]. Also, all variables included in the substantive model should be included in the imputation model [168]. The outcome variables must also be included, to ensure that the imputed covariate values have the correct association with the outcome [169, 170].

Incompatibility between imputation and substantive models

When substantive models include non-linear covariate effects, interactions, or are themselves non-linear (as hazard models), default imputation model do not assure compatibility between the substantive and the imputation model. To present this issue the notation presented in [171] was followed. Lets consider that the interest lies in modelling a fully observed outcome Y dependent on a single partially observed covariate X and a set of fully observed covariates $\mathbf{Z} = (Z_1, \cdots, Z_q)$. The substantive model is characterised by the function $f(Y|X, Z, \psi)$, where $\psi \in \Psi$ represents the substantive model parameters. Assuming MAR, the imputation model is characterised by $f(X|Z, Y, \omega)$, where $\omega \in \Omega$ represents the imputation model parameters. The imputation model is said to be compatible with the substantive model if there exists a joint model $g(Y, X|Z, \theta)$, $\theta \in \Theta$ and surjective maps $t_1 : \Theta \to \Omega$ and $t_2 : \Theta \to \Psi$ such that:

- for $\omega \in \Omega$, and $\theta \in t_1^{-1}(\omega) = \{\theta : t_1(\theta) = \omega\}$, $f(X|Z, Y, \omega) = g(X|Z, Y, \theta)$;
- for $\psi \in \Psi$, and $\theta \in t_2^{-1}(\psi) = \{\psi : t_2(\theta) = \psi\}$, $f(Y|X, Z, \psi) = g(Y|X, Z, \theta)$.

The two models are said to be semi-compatible if, by setting certain parameters in one or both models to zero, they can be made compatible. Incompatibility between the imputation and substantive models implies that, assuming the substantive model is correctly specified, the imputation model is mis-specified.

Substantive model compatible fully conditional specification

To overcome the problem of incompatibility between imputation and substantive models, Bartlett and colleagues [34] developed an algorithm based on rejection sampling that has been named Substantive model compatible-Fully conditional specification (SMC-FCS). Bartlett starts by noting that to specify an imputation model that is compatible with the substantive model:

$$\begin{aligned} f(X_j|X_{-j},Z,Y) &= \frac{f(Y,X_j,X_{-j},Z)}{f(Y,X_{-j},Z)} \\ &\alpha f(Y|X,Z)f(X_j|X_{-j},Z) \end{aligned} \tag{2.62}$$

So, in the algorithm SMC-FCS, a model $f(X_j|X_{-j},Z,\phi_j)$ must be specified, together with noninformative priors $f(\phi_j)$. Given values of ψ and ϕ, the missing values in X_j are imputed from the density proportional to:

$$f(Y|X,Z,\psi)f(X_j|X_{-j},Z,\phi_j) \tag{2.63}$$

Since generally this density does not belong to a standard parametric family, drawing samples from it is non-trivial [171]. Bartlett and Morris proposed a rejection sampling procedure that involves repeatedly drawing samples from a candidate distribution $f(X_j|X_{-j},Z,\phi_j)$ until the drawn value X_j satisfies the condition:

$$U \leq \frac{f(Y|X_j^*,X_{-j},Z,\psi)}{c(Y,X_{-j},Z,\psi)}, \tag{2.64}$$

where U is a random draw from an uniform distribution on (0,1) and $c(Y,X_{-j},Z,\psi)$ is an upper bound (in X_j) of $f(Y|X_j,X_{-j},Z,\psi_j)$ that does not involve X_j.

Diagnostics for imputations

Diagnostic techniques of the imputation procedures tend to be neglected. Most of the imputation methods rely on the MAR assumption. Since this assumption is untestable from observed data, this control tends to be ignored [172]. Diagnostic techniques can be characterised as *external*, if they involve outside knowledge, or *internal* if they depend on the observed data and the modelling process.

The first recommended approach is to check the imputed values. This can be done by comparing the distributions of the observed and imputed values graphically by means of histograms, boxplots, density curves, cumulative distribution plot, quantile-quantile plots or by means of descriptive measures [173]. The imputed distribution must not be, necessarily, similar to the observed values under MAR but extreme deviations should be assessed. Also, the plausibility of the imputed values should be checked given subject matter knowledge. Other possible diagnostic technique is checking the goodness-of-fit of the imputation models.

Sensitivity analysis

Multiple imputation relies on the MAR assumption. However, the true mechanism that led to data being missing can be MNAR, i.e. the probability of a variable being missing can be dependent on the missing value itself. The conclusions obtained under the MAR assumption should thus be tested to check how much sensitive they are to plausible deviations from that assumption.

There are essentially two approaches for performing sensitivity analysis: pattern mixture models and selection models [151]. In pattern mixture models, first the imputation model is fitted using the complete cases. Then, the values of the imputation model parameters are changed in order to reflect the deviations from MAR assumed as plausible (based on external knowledge). Third, the missing values are imputed using the imputation model with modified parameter values. Fourth, the substantive model is fitted to each completed dataset and the results combined using Rubin's rules [29].

The alternative approach is to use selection models. In this framework there are two models. The model of interest and a model for the chance of observations being missing. Both need to be fitted jointly.

2.8.7 Multiple imputation in survival analysis

Survival data is characterised by the fact that the variable of interest (time to event) is not observed for all individuals. This fact occurs due to censoring, as discussed above, and can be regarded as a missing data problem. However, the subject of this thesis does not focus on that situation but rather on the situation of having missing data on the covariates, i.e. in the variables that influence survival time. Multiple imputation was first applied to deal with missing covariates in survival analysis by van Buuren and colleagues in 1999 [162].

In covariate missing data problems, it is consensual that the outcome should be included in the imputation model. However, different ways of including the survival outcome can be found in the literature: D and T [174]; D and $log(T)$ [175, 176]; D, T and $log(T)$ [162] (D - censoring indicator; T - survival time). In 2009, White and Royston [177] recommended the inclusion of the cumulative baseline hazard ($H_0(t)$) besides the censor indicator. In spite of these recommendations, it is possible to find recent studies that did not use any outcome in the imputation model [178].

In 2015, Bartlett and colleagues implemented the SMC-FCS algorithm described above to Cox proportional hazards model [34]. In relative survival framework, Giorgi and colleagues were the first to introduce MI to deal with missing values on the covariates [31]. They performed a simulation study for different missing mechanisms and different missingness proportions, where they considered the proportional hazards assumption and that missing values occurred only on binary covariates. The MICE algorithm was used for imputation but the variables used in the imputation model were not clearly specified. They conclude that MICE performs well in estimating the hazard ratios and the baseline hazard function when the missing mechanism is MAR conditionally on the vital status. Multiple imputation has then been applied in the context of excess hazard estimation by different authors [107, 179] but without giving much detail in its application. Nur and colleagues published in 2010 a tutorial on handling missing data in relative survival analysis [27]. There they advocate the use of the MICE algorithm, the inclusion of the vital status and follow-up time in the imputation models as well as all the variables that are included in the analysis model, together with any interactions. Also, they recommend including as many predictors as possible in the imputation model to make the assumption of MAR more reasonable. In that work, the excess hazard was modelled using a generalised linear model with Poisson error where piecewise constant hazards are assumed.

Since then, the application of MI to deal with missing values in covariates can be found in some relative survival analysis literature [180, 181, 182, 183, 184].

Recently, Falcaro and colleagues evaluated the use of MI in the context of net survival problems with missing information on categorical covariates (stage of disease at diagnosis), first in the excess hazard modelling using flexible parametric proportional hazards models [32] and then in the non-parametric net survival estimation [33]. In the first study, the performance of different imputation strategies was evaluated using simulated incomplete datasets. The results obtained suggested that a multinomial logistic imputation model for stage should be used, instead of an ordered logistic model, and that the Nelson-Aalen cumulative hazard estimate and the event indicator should be included in the imputation models, as had been already suggested by White and Royston in the context of the Cox model [177]. The study had the limitation of considering only predictors whose effect was assumed to be constant over time. In the second, a resampling study was performed to evaluate the non-parametric estimation of net survival using the PP estimator after MI. Low bias and acceptable coverage rates for stage-specific net survivals were obtained after combining the estimates obtained for each completed dataset.

Although all the results found in these studies point MI as a valuable approach for dealing with covariate missing data in net survival and excess hazard modelling problems, the issue of compatibility between the imputation and substantive models still lacks some research when the substantive model is an excess hazard model.

Chapter 3

Studies

3.1 Study I: Age-standardised net survival estimation

Estimation of age-standardised net survival with sparse data: taking advantage of regression models
Luis Antunes, Denisa Mendonça, Aurélien Belot, Hadrien Charvat and Bernard Rachet

In this section, a study on the evaluation of methods to estimate age-standardised net survival is presented. Age-standardisation of net survival is common practice on research studies using population-based data to compare outcomes between different periods or regions. The direct standardisation method implies the estimation of age group-specific survival estimates. When dealing with small samples, common situation for cancers with low incidence, in small regions and/or for narrow time periods, the number of cases by age group can be insufficient to obtain reliable estimates.

Using a model-based approach to estimate net survival, i.e. fitting an excess hazard model to the data and then predicting survival from it, the performance of two different methods to obtain predictions were compared using a simulation study. The *classical* model-based approach consists in using the sample age structure to obtain age group specific predictions. This is dependent on the availability of cases and its age distribution. Alternatively the predictions were done for single reference ages in each age group.

This eliminates the variability induced by variations in the samples age structure. The formulas for the calculation of the variance and respective confidence intervals of the model-based age-standardised net survival were derived.

In the simulation study it was observed that the model fitting is the problematic step of the process. A high variability in age-specific model predictions between samples was obtained bringing up the difficulties in fitting more complex models to small datasets.

Next, the resulting manuscript of this study is presented.

Estimation of age-standardised net survival with sparse data: taking advantage of regression models

Luis Antunes [1,2,3], Denisa Mendonça [3,4], Aurélien Belot [5], Hadrien Charvat [6], and Bernard Rachet [5]

[1] Cancer Epidemiology Group, IPO Porto Research Centre (CI-IPOP), Portuguese Oncology Institute (IPO-P), Porto, Portugal

[2] Faculty of Sciences, University of Porto, Porto, Portugal

[3] EPIUnit, Institute of Public Health, University of Porto, Porto, Portugal

[4] Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal

[5] Cancer Survival Group, London School of Hygiene and Tropical Medicine, London, United Kingdom

[6] Section of Cancer Surveillance, International Agency for Research on Cancer, Lyon, France

Address for correspondence: Luis Antunes, Department of Epidemiology, Portuguese Oncology Institute (IPO-P), Rua Dr. António Bernardino de Almeida, 4200-072, Porto, Portugal.

E-mail: `luis.antunes@ipoporto.min-saude.pt`.

Phone: (+351) 225 084 000.

Fax: (+351) 225 084 001.

Abstract: Cancer survival comparisons between different regions or periods are performed using an age-standardised net survival measure to account for heterogeneity in population age structures, age being the main confounder. Direct age-standardisation requires the estimation of survival by age group. When comparing less frequent tumours or smaller populations, often data are sparse, turning the estimation of age-standardised survival impossible or unreliable. The aim of this study was to evaluate the performance of the conventional parametric and non-parametric approaches to estimate this measure comparing with an alternative approach for standardisation, where the age group-specific survival estimates are independent from the sample age structure. A simulation study was performed to compare the different approaches. The alternative proposed method presented similar or higher performance results as the conventional approaches. It was shown to be a valid alternative mainly when age group specific estimates are not possible to estimate. As illustration, the estimation of age-standardised net survival from vagina cancer for a set of patients diagnosed in the North region of Portugal was performed using the different methods. While using the proposed method allowed the estimation of age-standardised net survival for all the individual years of diagnosis, using the conventional model-based and the non-parametric approach this was only possible in less than half of those subgroups.

Key words: net survival, age standardisation, excess hazard model, sparse data

1 Introduction

Population-based cancer survival analysis is of major importance in the evaluation of cancer care practices provided to populations. The measure of survival used in these evaluations should be related to mortality from the cancer under study and not to all-cause mortality. One of the key indicators is net survival, i.e., the survival that would be observed in the absence of other causes of death. Net survival is estimated from observed mortality after taking into account background (all-causes) mortality obtained from population life tables.

Net survival is, for most cancers, age-dependent. International comparison of net survival probabilities should thus take into account differences in patient´s population age structure. This is usually achieved through direct age-standardisation using a common age-distribution set of standards such as the International Cancer Survival Standards (Corazziari et al. (2004)). The direct age-standardisation implies the estimation of net survival by age group. In some situations, the extreme age groups (youngest or oldest, depending on the cancer) are sparse and as a consequence, the age group-specific net survival may not be estimated because of no observations or no observations remaining after a short follow-up time. Cancer survival studies (e.g., EUROCARE (De Angelis et al. (2014)), CONCORD (Allemani et al. (2015, 2018)) face this issue when estimating net survival for small countries and/or for rarer cancers. When age-group specific estimates are not possible or are unreliable, contiguous age-groups are aggregated and the common estimate is assigned to both age groups. Otherwise, only unstandardised estimates are presented. Using non-parametric estimators this means calculating a survival estimate for the entire sample without stratifying by age groups. In both situations, the comparison with other age-standardised estimates is less reliable.

An alternative approach of age-standardising survival has been proposed by Brenner in 2004 in which weights are attributed directly to individual patients instead of weighting age-group specific survival estimates (Brenner et al. (2004)). The unreliability of estimates within age-groups with sparse data remained though a concern using this alternative method (Gondos et al. (2006)).

Net survival can be estimated using the non-parametric Pohar-Perme estimator (Perme et al. (2012)) or using a modelling approach (Danieli et al. (2012)). If the model is correctly specified both methods should produce asymptotically similar estimates. When age is considered as a continuous variable and the excess hazard is modelled with flexible functions (e.g. splines), net survival of each individual can be thinly predicted for any time since diagnosis. The net survival of a given age group is obtained as the mean of the individual net survival predictions of the subjects in this age group. Although a flexible modelling approach is used, estimates of age group-specific net survival depend on the observed number of subjects in each age group at the time of diagnosis, as well as their observed age-distribution in each age group. When the data are sparse, this will lead to unstable net survival estimates even if the model allows to smoothly predict exact individual net survivals. Furthermore, it is possible that some age groups have no observations making it impossible to estimate age group specific survival. Estimates given by the non-parametric Pohar-Perme estimator are also very unstable or impossible in these situations.

The main aim of this study was to evaluate and compare the methods classically used for the estimation of age-standardised net survival (model-based and non-parametric) with an alternative proposed approach when data are sparse. In this alternative, besides the conventional external standardisation using the Corazziari weights, net survival within each age group is estimated for a reference age corresponding to a complementary external standardisation.

The manuscript is organized as follows: in Section 2, the statistical methods used to estimate age-standardised net survival, both parametrically and non-parametrically, are summarised. Also, the alternative approach for estimating age-standardised net survival is described. The simulation study and its results are presented in Section 3 followed by a real-world illustration in Section 4. Section 5 concludes the manuscript with a discussion.

2 Estimation of age-standardised net survival

2.1 Net survival estimation

Net survival is defined as the survival that would be observed in the hypothetical situation that the disease is the only cause of death possible. In the relative survival setting it is assumed that the overall hazard for patient i ($\lambda_{O_i}(t)$) can be decomposed in two additive components:

$$\lambda_{O_i}(t) = \lambda_{P_i}(t) + \lambda_{E_i}(t) \tag{2.1}$$

where $\lambda_{P_i}(t)$, the expected hazard, is given by the general population mortality assuming this to be a reasonable approximation of other causes of mortality and $\lambda_{E_i}(t)$ is the excess hazard due to the disease in study. The population mortality ($\lambda_{P_i}(t)$) is obtained from life tables, usually made available by the National Statistics Offices, stratified by relevant demographic variables (e.g., sex, age). The survival obtained by exponentiating minus the integral of the excess hazard is our measure of interest, the net survival. This survival can be estimated non-parametrically or using model-based approaches.

2.1.1 Pohar-Perme estimator

A consistent and unbiased non-parametric estimator of net survival, the so-called Pohar-Perme (PP) estimator, was proposed in 2012 (Perme et al. (2012)). This estimator accounts for the bias due to informative censoring by weighting the individuals still contributing to the parameter estimation. The weights correspond to the inverse of the individual-specific expected survival probabilities provided by the general population life tables. In this way the early drop off from the sample of the patients with higher expected mortality is compensated through a higher contribution (weight) to the group survival. Let $N_i(t) = I(T_i \leq t, T_i \leq C_i)$ and $Y_i(t) = I(T_i \geq t, C_i \geq t)$ denote the counting process and the at-risk process for each individual in the sample, where T_i denotes the time to death from any cause and C_i the time to censoring for patient i. The PP estimator weights these two processes using the inverse of the population survival probability: $N_i^w(t) = N_i(t)/S_{Pi}(t)$ and $Y_i^w(t) = Y_i(t)/S_{Pi}(t)$, thus providing an estimate of the cumulative excess hazard:

$$\hat{\Lambda}_E(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)}, \qquad (2.2)$$

where the subscript i represents each patient, n is the total number of patients in the group, $N^w(t) = \sum N_i^w(t)$ and $Y^w(t) = \sum Y_i^w(t)$. The net survival for the group of patients is then obtained by $\hat{NS}(t) = e^{-\hat{\Lambda}_E(t)}$.

2.1.2 Model-based estimation

Net survival can also be estimated from a multivariable regression model (Perme et al. (2012); Danieli et al. (2012)). The excess hazard function is modelled as a function of a set of covariates, including at least the demographic variables which the estimation of the expected mortality is stratified on. Considering a flexible parametric model for the excess hazard function, where the baseline excess hazard is flexibly modelled using splines, non-linear and time-dependent effects of covariates are allowed and can be modelled using B-spline function (Giorgi et al. (2003); Remontet et al. (2007); Charvat et al. (2016)). The model can be written as:

$$\log[\lambda_E(t,x)] = \log[\lambda_0(t)] + g(\mathbf{X}) + \mathbf{X} \cdot h(t), \tag{2.3}$$

where $\log[\lambda_0]$ and h are B-spline functions and g can be a linear or non-linear function of the covariates $\mathbf{X}$. We used the implementation of this model following the work of Charvat and colleagues (Charvat et al. (2016)). The net survival of a patient can be predicted from the model integrating the excess hazard:

$$S_i(t,x) = exp\left\{-\int_0^t \lambda_{E_i}(u,x)du\right\} \tag{2.4}$$

The net survival in each age group is obtained by averaging the individual survivals of the patients within each group.

$$NS_j(t,x) = \frac{1}{n_j}\sum_{i=1}^{n_j} S_i(t,x) \tag{2.5}$$

where n_j is the number of patients in age group j. We will further refer this method as the *classical* approach. The net survival in each age group thus depends not only on the individual survival but, being survival age-dependent, depends also on the age distribution of patients within each group.

2.2 Age-standardised net survival

The aim of standardising survival probabilities according to an age distribution is to make the estimation of this quantity comparable between two populations with different age-structures, age being implicitly considered the main confounder. If two populations observed in two different countries are exposed to the same age-specific survival probabilities, the age-standardised survival obtained from each country should be equal. A way to do so is to standardise on a user-defined external distribution: $ASNS(t) = \int_z S(t|z)dG(z)$, where G represents that external distribution. This standard population is usually approximated by a discrete distribution. The age-standardised estimate is thus given by a weighted average of age group-specific net survival estimates ($NS_j(t)$):

$$ASNS(t) = \sum_{j=1}^{k} w_j \cdot NS_j(t) \tag{2.6}$$

where k is the number of age groups considered and w_j are the respective weights ($\sum w_j = 1$). In population-based cancer research, to allow comparability between different studies, common weights are used as defined by the International Cancer Survival Standard (Corazziari et al. (2004)). In this standard, five age groups are considered (for most cancers: 15-44, 45-54, 55-64, 65-74, 75+). For cancers with increasing incidence by age (most cancer sites) the standard weights are 0.07, 0.12, 0.23, 0.29, 0.29, respectively for the age groups defined above.

The variance for the non-parametric estimate of age-standardised survival can be given by:

$$VAR(ASNS) = \sum_{j=1}^{k} w_j^2 \cdot SE_{S_j}^2 \tag{2.7}$$

assuming independence between the age group-specific survival estimates, and $SE_{S_j}^2$ being the variance of the net survival for age group j.

The estimation of the variance of the model-based age-standardised net survival needs to account for the correlation between two individuals' net survival prediction since they are derived from the same regression coefficients. The derivation of an approximate formula to estimate this variance using the delta method is presented in Supplementary Material S1.

2.3 Alternative approach for age-standardised net survival estimation

Although a flexible modelling approach is used in the model-based method described above, estimates of age group-specific net survival depend on the number of subjects in each age group at the time of diagnosis, as well as their age distribution in each age group. The comparability between different populations can be compromised due to this dependence of the own sample age structure. Also, in the presence of sparse data, some age groups may contain no observation turning it impossible to estimate age-group-specific survival in those age groups. To circumvent the variability induced by sample variations in the age distribution and the impossibility of estimating age-standardised net survival estimates, we propose an alternative approach for the estimation of a survival measure that allows comparison between different populations or time periods, even in these situations of sparse data. Instead of averaging model-based individual net survival predictions within each age group, we propose to estimate survival for each group at a pre-specified reference age value. We further refer to this approach as the *alternative* approach. Having the fitted model, the predicted net survival for each age group would no longer be dependent on the age distribution neither on the existence of observations within each group since it is calculated at a reference age, externally defined, instead of being given by the average of the individual net survival predictions. This approach uses thus an external standardisation to estimate survival within each age group. Considering that the excess hazard is only dependent on age, the net survival for each age group j is given by:

$$NS_j^*(t) = exp\left\{-\int_0^t \lambda_E(u, age_{ref_j})du\right\} \tag{2.8}$$

The reference age can be an international standard to be defined specific for each cancer site.

3 Simulation study

3.1 Study description

A simulation study was performed in order to compare the performances of the methods described for estimating age-standardised net survival. Data were generated based on real datasets. These were extracted from the North Region Cancer Registry of Portugal (RORENO) database, a population-based registry which covered until 2010 a population of around 3.2 million inhabitants. Survival was considered as being dependent solely on age at diagnosis and year of diagnosis. To test the performance of the methods in a broad range of situations, similarly to what is found in real-world data, three scenarios of different complexity were considered, using information on three different cancer sites as starting datasets. The three scenarios considered in the simulation study were:

- Scenario A (*Stomach cancer*): Non-linear and time-dependent effect of age, no effect of year of diagnosis;
- Scenario B (*Breast cancer*): Non-linear and time-dependent effect of age, linear and proportional effect of year of diagnosis;
- Scenario C (*Colon cancer*): Non-linear and time-dependent effect of age, linear and time-dependent effect of year of diagnosis plus interaction age × year.

Generation of time to cancer related death

The covariates used in this simulation study were age and year of diagnosis. The proportion of patients in each age group was obtained from the real datasets. Within each age group, age was drawn from a uniform distribution. Year of diagnosis was generated from a uniform distribution. For scenarios A and C only male patients were considered and for scenario B (breast cancer), only female patients. A period of ten years of diagnosis was considered: 2001-2010.

In order to avoid using exactly the same model in the data generation and in the data fitting steps, an excess cumulative hazard model using fractional polynomials for the baseline was used in the data generation step (Lambert et al. (2005)). This models was fitted to each real dataset and the estimated parameters were used to generate the survival times from cancer.

Generation of time to other causes death

Times to death from other causes were generated using a piecewise exponential distribution and following the same scheme used in similar studies (Rutherford et al. (2012); Charvat et al. (2016)).

The final observed survival time was taken as the minimum between time to death from cancer, time to death from other causes or a censoring time due to the end of follow-up (pre-set on the 31st December 2015). One million observations were generated for each scenario from which 1000 samples of two different sizes (n=200 and n=2000) were randomly selected. The number of simulations chosen allows the estimation of the quantity of interest with an error margin lower than 1% (Burton et al. (2006)).

Modelling approaches

For the estimation of age-standardised net survival using model-based approaches, a general flexible parametric model for the excess hazard function was considered (Charvat et al. (2016)):

$$\log[\lambda_E(age, year, t)] = \log[\lambda_0(t)] + s(age) + year + age \cdot g(t) + year \cdot g(t) + age \times year$$

A B-spline function of third degree with two fixed internal knots for the log-baseline hazard was considered. Non-linear effect of age was considered using a truncated power basis spline with one knot. Time-dependent effects of age and year of diagnosis modelled by introducing interaction terms between these variables and the same Bspline as the one used to model the logarithm of the baseline hazard (g). Interaction between age and year of diagnosis was also considered.

Two strategies were used for model selection. The first, considered that the population model best describing the data to be fitted is known. Since with real datasets, this model is unknown, we tested also a second model selection strategy where for each sample the most parsimonious model was chosen for each sample:

1. *FixedMod*: Fixed type of model (according to specific scenario) for all samples (only coefficients re-estimated);

2. *BestMod*: Choose 'best' model for each sample based on a backward selection algorithm (described in Supplementary Material S2);

For each sample, age-standardised net survival at 5 years was estimated for the all period (2001-2010) and by year of diagnosis (using the model fitted to the all period). Model-based predictions were calculated by averaging the individuals net survival in each age group (*classical* - equation 2.5) and by calculating survival for a reference age in each age group (*alternative* - equation 2.8). Furthermore, the non-parametric estimates were also calculated using the Pohar-Perme (PP) estimator. Whenever the model-based predictions using the *classical* method, or the PP estimator, were not possible for any age group, the unstandardised net survival was estimated instead, i.e. estimated for the full sample without stratifying by age group.

To evaluate and compare the performance of the several approaches in estimating the age-standardised net survival (ASNS), the estimates obtained in each scenario were compared with the population values. These true (population) values were obtained from the full dataset generated. Two population values were considered. One for the evaluation of the performance of the non-parametric approach and of the *classical* model-based approach, and another for the evaluation of the *alternative* approach. In the first case, each age group-specific population value was calculated using an internal standardisation, i.e. considering the population age structure. In the second case, age group-specific values were obtained by predicting survival at a reference ages, i.e. using an external standardisation. In this simulation study, in the absence of defined reference standards, the reference ages have been considered as being the average age within each age group of the population where samples were drawn from. The following performance measures were calculated for each situation: (i) the bias (mean difference between estimates and true value); (ii) the percentage bias (bias divided by the true value times 100); (iii) the empirical coverage probability (CP) (proportion of 95% estimated confidence intervals that included the true value of the ASNS; (iv) the root mean square error (RMSE) (square root of the average of the squared differences between the estimated values and the true value). According to Burton and colleagues (Burton et al. (2006)), for 1000 simulations, between 936 and 964 of the confidence intervals should include the true value.

Before running the simulations with the enumerated modelling strategies, we first evaluated the variability solely induced by variations on the sample age distribution. Net survival as function of age was assumed to be known (the model fitted to the full dataset was used). Random samples were drawn from the full dataset and age-standardised net survival was calculated using the *classical* approach. In this setting, the predictions obtained with the *alternative* approach were constant since the model used for predictions was always the same for all samples.

In the data generation step, the statistical software STATA (StataCorp (2011)) was used with the packages *survsim* and *stpm2*. In the data analysis step, R software was used (R Core Team (2017)), namely, the package *mexhaz* (Charvat and Belot (2017)) for excess hazard model fitting and the package *relsurv* (Maja Pohar Perme (2016)) for the Pohar-Perme estimator.

3.2 Results

3.2.1 Constant model

At first, we aimed to evaluate the variability in ASNS estimates induced by variations on the sample age distribution. Due to lack of observations in at least one age group when considering only one year of diagnosis (2001), it was not possible to calculate age-standardised net survival in 29.7%, 11.9% and 55.9% of the samples for scenarios A, B and C, respectively. In these situations, the ASNS was replaced by the unstandardised measure. Figure 1 shows the variability of the 1000 ASNS estimates for the all period of diagnosis and the one obtained when estimating only for one specific year (2001). In scenarios B and C the median of the survival results for 2001/10 and 2001 are different since survival was considered year dependent. These results show that even using always the same model, there is a considerable variability caused by the variation in the samples age distribution.

3.2.2 Model fitted to each sample

Next, we fitted a model for each of the thousand samples randomly drawn for the population. The two strategies for choosing the model, described above, were used. Net survival by 1-year age-group was predicted using the retained model fitted to each sample. The 1000 predicted net survivals as function of age, their mean and the comparison with the true values are presented in figure 2 for scenarios A, B and C. The graphs on the left hand side of the figure correspond to the *FixedMod* model fitting strategy and on the right hand side to the *BestMod* strategy. In scenario A, survival does not depend on year of diagnosis. For scenarios B and C, where it does, the presented results correspond to the reference year of diagnosis (2005). For sample size of 200, a large variability in the age-specific predictions was observed especially in younger ages (below 40) and older ages (above 80). For scenarios A (no effect of year) and B (proportional effect of year), using the strategy *FixedMod*, a good agreement between the mean of the simulations and the true value was observed except for very young ages where survival tended to be slightly underestimated. In scenario C (time-dependent effects of age and year of diagnosis plus interaction age × year) there was a small but systematic bias between the estimated values and the true value.

When not assuming that the type of model is known (*BestMod*), the agreement between the mean of simulated values and the true value was lower than when using the *FixedMod* strategy. However, the interquartile range of the model-based predictions was much narrower in the age extremes. Due to small sample sizes, the simplest model (linear effect of covariates and proportional hazard assumption) was chosen by the algorithm most of the times: 81.0% (A), 65.6% (B) and 51.4% (C). The results obtained for samples of size 2000 are presented in the Supplementary Material S3. As expected, for this sample size, the variability was much lower and the agreement between the mean and the true values for the *BestMod* strategy increased.

3.2.3 Comparison of ASNS estimation approaches

Finally, we aimed at comparing the performance of the *classical* and the *alternative* approaches for each of the model selection strategies. For the classical model-based approach, whenever there were no observations in a specific age group, the unstandardised net survival was estimated instead. Using the non-parametric estimator, besides those situations, age-standardised net survival was not possible to calculate also when there were an insufficient number of events to estimate net survival by age group at 5-years. Table 1 presents the percentage of samples for which it was possible to estimate the age-standardised measure. The lowest values were obtained with the non-parametric estimator when trying to estimate for specific years of diagnosis (10.3%, 66.7%, 18.5% for scenarios A, B and C, respectively).

For each of the situations studied, the performance measures obtained, namely, bias, percentage bias, empirical coverage probability and RMSE, are presented in Table 2. Model convergence was attained for all samples, except on scenario B (*FixedMod*) where convergence was attained for 98.6% of the samples. In general, bias in the estimation of age-standardised net survival was lower when the type model that best described the data was assumed to be known (*FixedMod*). When estimating net survival for the all period of diagnosis, the bias obtained with the *classical* and the *alternative* approaches were similar. But, when estimating for smaller sets, that is, for a specific year of diagnosis, the bias achieved with the *alternative* approach was in general lower than with the *classical* one. The empirical coverage probability was inside the expected range for scenarios A and C, except when using the *classical* approach in situation B-*BestMod*, where it was 91.9%. For scenario B, all values were below the expected, ranging from 88.9% to 91.5%. There were no major differences between the values of RMSE obtained for the *classical* and *alternative* approaches for all scenarios and model choosing strategies. The values estimated for 2001 with the PP estimator presented in general higher bias and higher RMSE than the model-based results, but the empirical coverages were similar. These results for samples of $n = 2000$ are presented as Supplementary Material S3.

4 Trends in age-standardised net survival from vagina cancer

To illustrate the application of the proposed method and compare it with the established methods, the estimation of age-standardised net survival was performed for a real dataset of vagina cancer patients diagnosed in the North region of Portugal. This is a rare cancer topography with less than 15 new cases per year diagnosed in the region considered.

4.1 Data and methods description

The data were extracted from RORENO. All patients diagnosed with vagina cancer (ICD10: C52) in the period 2001-2010 and followed-up until the end of 2015, with residence in the area covered by the registry were considered for analysis. A total of 122 cases were considered eligible. After excluding patients with unknown age or unknown survival time, 116 patients were included in the analysis. The median age at diagnosis (P25-P75) was 68 (56-77).

The objective of the analysis was to estimate age-standardised net survival by year of diagnosis. Net survival was estimated by age group, considering the same five age groups described in the simulation study. Age-standardisation was calculated as the weighted sum of the age group specific survivals using the ICSS weights described in section 2.2.

A model building strategy as described in Supplementary Material S2 was followed to fit a model to the ten-year period, considering age and year of diagnosis as covariates. Net survival was then predicted for each individual in the sample using the fitted model. The age group and year of diagnosis specific net survival was estimated by averaging the individual predicted survival in each group. For each year of diagnosis, when there was at least one observation by age group, age-standardised net survival was calculated. Otherwise, only unstandardised net survival was calculated for those years. Furthermore, age-standardised net survival was estimated using the proposed alternative approach. The reference age considered for each age group was its mean age for the all ten years period. Non-parametric net survival estimates were also obtained using Pohar-Perme estimator. Whenever possible, age-standardised net survival was calculated.

4.2 Results

The final model obtained considered only linear effects of age and year of diagnosis and assumed proportional hazards for both variables. Using the *classical* approach it was only possible to calculate age-standardised net survival for four of the ten different years of diagnosis. For the years for which it was possible, the difference between the values obtained by this and by the *alternative* approaches was minimal. Estimating year-specific age-standardised net survival using the non-parametric estimator was only possible for one of the years. On the contrary, the *alternative* method allowed the estimation of survival for all individual years, showing a smooth trend in survival along the years. The unstandardised net survival was also possible to estimate for all years but showing an unstable trend overtime (Figure 3).

5 Discussion

When comparing net survival between two populations, it is desirable to have a measure that produces the same estimate if both populations have the same age-specific survival. Since age distribution between populations can vary, standardized measures are used when performing this type of comparisons. However, the commonly used standardisation methods do not guarantee the correct comparability of survival between populations. Standardisation is calculated by weight-averaging age group-specific estimates of survival. The conventional non-parametric and model-based approaches for estimating those are dependent on the sample age structure within each age group. This dependence is minimal if age groups are narrow but can have a non-negligible impact when using broader age classes as the five age groups usually used in population-based cancer survival analysis. Also, with sparse data, direct standardisation can be difficult or impossible due to low number of cases/events especially in the extreme age groups. The aim of this study was to compare the conventional age standardisation methods with a proposed alternative approach where age group-specific survival is estimated using external weights.

We first studied the performance of the proposed approach in the unrealistic assumption that the model that best described the full population was known. Forcing a relatively complex model to be fitted to small datasets led to unstable model-based predictions, especially in younger ages (below 40) and older ages (above 80). Allowing a different model to be selected for each sample based on a backward elimination algorithm retaining only the significant effects, resulted in narrower interquartile ranges of age-specific survival predictions. This was mainly due to the selection in a considerable proportion of cases of a simple proportional hazards model with linear effect of age. The sharp asymmetrical distribution of age-specific predicted survivals in extreme age groups, especially in scenario B, led to higher deviation between the simulated mean and the true value. Using samples of larger size, the variability in the prediction was much smaller as well as the difference between the simulated mean and true (population) values.

The proposed alternative approach, using an external standardisation complementary to the conventional external standardisation using the Corazziari weights aimed at solving two issues of the *classical* approach. Lack of observations in specific age groups and differences in age-standardised net survival between populations, with the same survival function, induced by variations in age distribution. The simulation results shown that the variability of the results obtained with the *classical* and the *alternative* approaches did not differ much. This variability has two sources: variability of the model fitted to the data; variability due to variations in age distribution. The *alternative* approach eliminates the second but not the first source. If the model fitted is then used to predict survival in subsets of data as by year of diagnosis in our study, the advantage in terms of variability using the *alternative* approach is more evident. Nonetheless, in all the situations the proposed method allows estimating survival for any subset even if there are none or low number of cases in specific age groups while using the *classical* approach this was possible in less than 50% of the situations and using the non-parametric estimator this percentage dropped to less than 20%.

The practical application presented further illustrated the advantages of the proposed approach. With this, it was possible to estimate an age-standardised measure for all years of diagnosis of interest while using the *classical* model-based method it was only possible for less than half of the years and using the Pohar-Perme just for 10%. To obtain more non-parametric estimates, we had to combine adjacent age groups.

In the proposed approach, age group-specific estimates were obtained from model-based predictions at reference ages. In the simulation study performed, these corresponded to the age population mean for each age group. In real-life applications, this mean would be unknown. Other alternative ways of estimating survival for each age group could be thought. Using the mean of each age class, an average of age-specific survival within age group or the use of any standard population externally defined are possible alternatives. In the SUDCAN study, Uhry and colleagues (Uhry et al. (2017)) used model-based predictions to estimate year- and country-specific age-standardised net survival. Prior to the usual external age-standardisation using the ICSS weights, the age class-specific survival was calculated by averaging the annual age-specific net survivals predicted from the model using the age weights within the age-class as observed over the entire data (country and site specific). By fixing the age structure, this approach had the purpose of preventing variations in ASNS estimates induced by age distribution changes over time within the age groups (especially in the older ones). Our approach sets a fixed age-structure also by considering a reference age for each age group, constant over time, suitable even when the number of observations in the entire data is sparse in extreme age groups. Also, we used five groups since it is the most common in international survival studies De Angelis et al. (2014); Allemani et al. (2015, 2018). The ICSS also defines weights for 5-year age groups. This would be suitable for model-based predictions but not for non-parametric estimation. The use of narrower age groups would increase the frequency of the issues we are trying to solve, namely, the low number of cases by age group. Model-based estimation can be an alternative to non-parametric estimators but have also its difficulties. A correctly specified model can be difficult to fit and also it relies in more assumptions than non-parametric estimation.

The issue of age-standardisation has been addressed in this study in the context of sparse data. The question of what is the best method to age-standardise an age dependent measure can however be addressed on a broader context of large samples. In either contexts, the standardisation should guarantee that the standardise measure does not depend on the internal age structure. Two aspects need to be considered when choosing a standardisation method: how to discretize the age distribution and how to estimate age group-specific survival. Age groups can be wider or narrower (eventually individual ages). Age group-specific survivals can be estimated non-parametrically or parametrically. In the last case, different alternatives for the point(s) for which survival is predicted can be considered. Further studies are needed to evaluate the different approaches in terms of their performance.

Acknowledgements

The authors would like to thank the North Region Cancer Registry of Portugal for the availability of the data used in this study. The authors would also like to acknowledge the contribution of Camille Maringe in the development of the simulation study.

References

Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X.-S., Bannon, F., Ahn, J. V., Johnson, C. J., Bonaventure, A., Marcos-Gragera, R., Stiller, C., Azevedo e Silva, G., Chen, W.-Q., Ogunbiyi, O. J., Rachet, B., Soeberg, M. J., You, H., Matsuda, T., Bielska-Lasota, M., Storm, H., Tucker, T. C., Coleman, M. P., and CONCORD Working Group (2015). 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). *The Lancet*, **385**(9972), 977–1010. ISSN 01406736. doi: 10.1016/S0140-6736(14)62038-9. URL http://www.ncbi.nlm.nih.gov/pubmed/25467588http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4588097http://linkinghub.elsevier.com/retrieve/pii/S0140673614620389.

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Nikšić, M., Bonaventure, A., Valkov, M., Johnson, C. J., Estève, J., Ogunbiyi, O. J., Azevedo E Silva, G., Chen, W.-Q., Eser, S., Engholm, G., Stiller, C. A., Monnereau, A., Woods, R. R., Visser, O., Lim, G. H., Aitken, J., Weir, H. K., Coleman, M. P., and CONCORD Working Group, S. (2018). Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet (London, England)*, **391**(10125), 1023–1075. ISSN 1474-547X. doi: 10.1016/S0140-6736(17)33326-3. URL http://www.ncbi.nlm.nih.gov/pubmed/29395269http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5879496.

Brenner, H., Gefeller, O., and Hakulinen, T. (2004). Period analysis for up-to-date' cancer survival data. *European Journal of Cancer*, **40**(3), 326–335. ISSN 09598049. doi: 10.1016/j.ejca.2003.10.013. URL http://www.sciencedirect.com/science/article/pii/S0959804903009237.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**(24), 4279–4292. ISSN 02776715. doi: 10.1002/sim.2673. URL http://www.ncbi.nlm.nih.gov/pubmed/16947139http://doi.wiley.com/10.1002/sim.2673.

Charvat, H. and Belot, A. (2017). *mexhaz: Mixed Effect Excess Hazard Models*. URL https://cran.r-project.org/package=mexhaz.

Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., and Belot, A. (2016). A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine*, **35**(2016), 3066–84. ISSN 02776715. doi: 10.1002/sim.6881. URL http://doi.wiley.com/10.1002/sim.6881.

Corazziari, I., Quinn, M., and Capocaccia, R. (2004). Standard cancer patient population for age standardising survival ratios. *European journal of cancer (Oxford, England : 1990)*, **40**(15), 2307–16. ISSN 0959-8049. doi: 10.1016/j.ejca.2004.07.002. URL http://www.sciencedirect.com/science/article/pii/S0959804904005283.

Danieli, C., Remontet, L., Bossard, N., Roche, L., and Belot, A. (2012). Estimating net survival: The importance of allowing for informative censoring. *Statistics in Medicine*, **31**(8), 775–786. ISSN 02776715. doi: 10.1002/sim.4464.

De Angelis, R., Sant, M., Coleman, M. P., Francisci, S., Baili, P., Pierannunzio, D., Trama, A., Visser, O., Brenner, H., Ardanaz, E., Bielska-Lasota, M., Engholm, G., Nennecke, A., Siesling, S., Berrino, F., and Capocaccia, R. (2014). Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE–5-a population-based study. *The Lancet. Oncology*, **15**(1), 23–34. ISSN 1474-5488. doi: 10.1016/S1470-2045(13)70546-1. URL http://www.thelancet.com/article/S1470204513705461/fulltext.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Esteve, J., Gouvernet, J., and Faivre, J. (2003). A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine*, **22**(17), 2767–2784. ISSN 0277-6715. doi: 10.1002/sim.1484. URL http://doi.wiley.com/10.1002/sim.1484.

Gondos, a., Parkin, D. M., Chokunonga, E., and Brenner, H. (2006). Calculating age-adjusted cancer survival estimates when age-specific data are sparse: an empirical evaluation of various methods. *British journal of cancer*, **94**(3), 450–454. ISSN 0007-0920. doi: 10.1038/sj.bjc.6602976.

Lambert, P. C., Smith, L. K., Jones, D. R., and Botha, J. L. (2005). Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine*, **24**(24), 3871–3885. ISSN 0277-6715. doi: 10.1002/sim.2399. URL http://www.ncbi.nlm.nih.gov/pubmed/16320260http://doi.wiley.com/10.1002/sim.2399.

Maja Pohar Perme (2016). *relsurv: Relative Survival.* URL https://cran.r-project.org/package=relsurv.

Perme, M. P., Stare, J., and Estève, J. (2012). On Estimation in Relative Survival. *Biometrics*, **68**(1), 113–120. ISSN 0006341X. doi: 10.1111/j.1541-0420.2011.01640.x.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/.

Remontet, L., Bossard, N., Belot, A., and Est, J. (2007). An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*, **26**(December 2005), 2214–2228. doi: 10.1002/sim.

Rutherford, M. J., Dickman, P. W., and Lambert, P. C. (2012). Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiology*, **36**(1), 16–21. ISSN 18777821. doi: 10.1016/j.canep.2011.05.010. URL http://dx.doi.org/10.1016/j.canep.2011.05.010.

StataCorp (2011). Stata Statistical Software: Release 12.

Uhry, Z., Bossard, N., Remontet, L., Iwaz, J., Roche, L., and GRELL EUROCARE-5 Working Group and the CENSUR Working Survival Group (2017). New insights into survival trend analyses in cancer population-based studies. *European Journal of Cancer Prevention*, **26**, S9–S15. ISSN 0959-8278. doi: 10.1097/CEJ.0000000000000301. URL http://www.ncbi.nlm.nih.gov/pubmed/28005600http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage{&}an=00008469-201701001-00003.

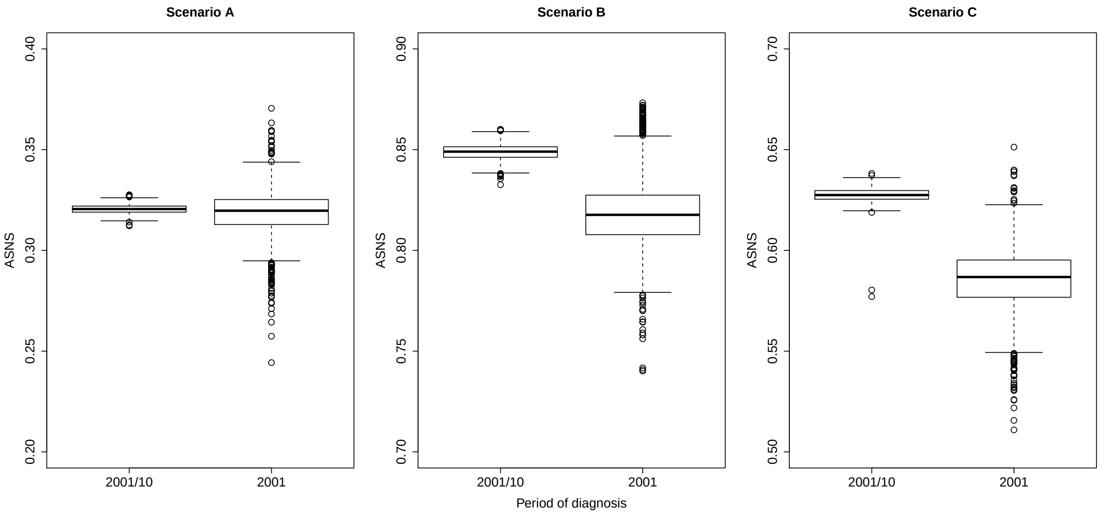

Figure 1: Age-standardised net survival by period of diagnosis (scenarios A, B, C) - predictions using population model.

Table 1: Percentage of samples for which it was possible to estimate age-standardised net survival ($n = 200$).

Scenarios	2001-2010		2001	
	classical	PP	*classical*	PP
A	100.0	99.9	70.8	10.3
B	100.0	100.0	88.5	66.7
C	100.0	99.9	45.4	18.5

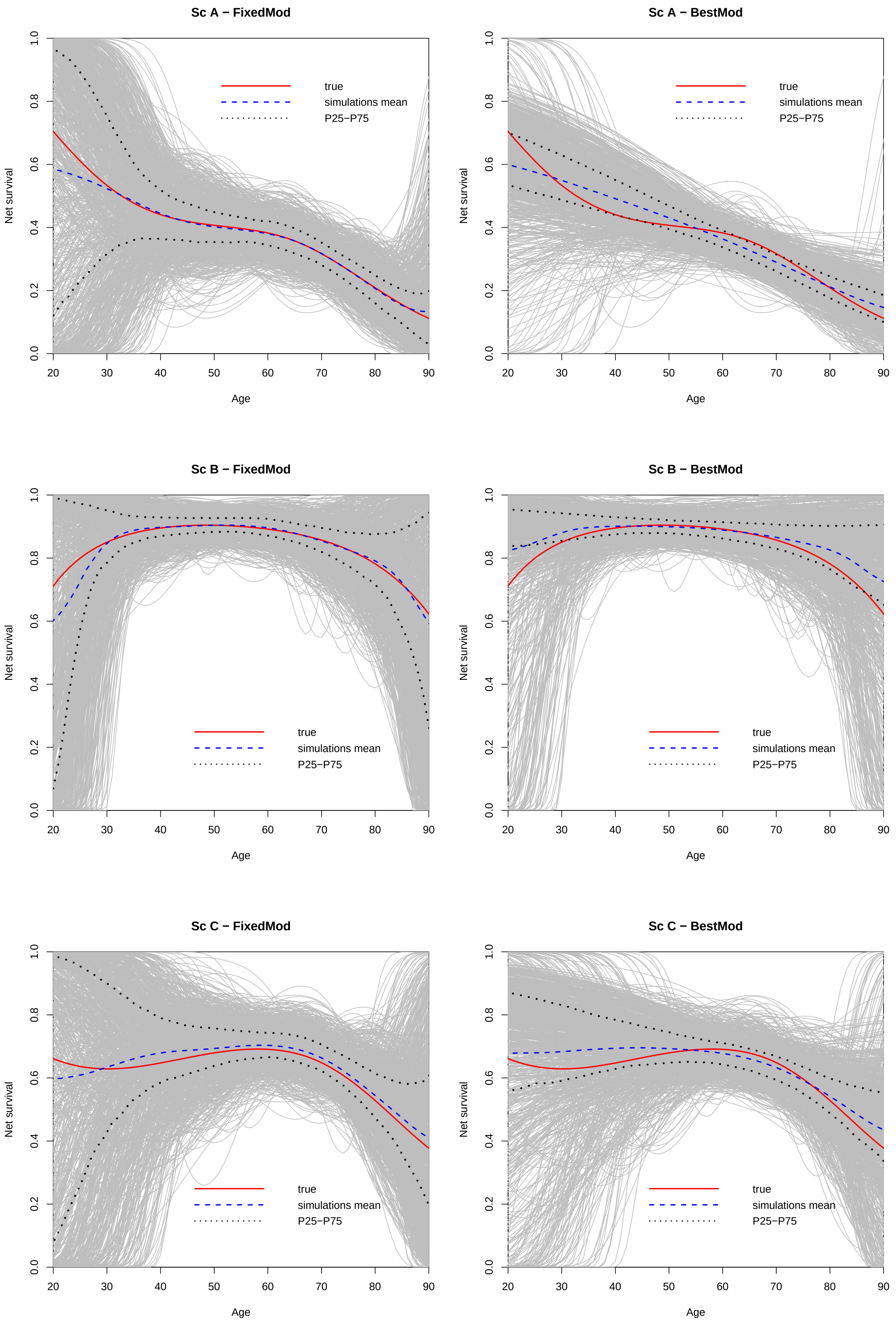

Figure 2: Predicted survival by age: a) Scenario A; b) Scenario B; c) Scenario C. Model fitted to samples of $n = 200$

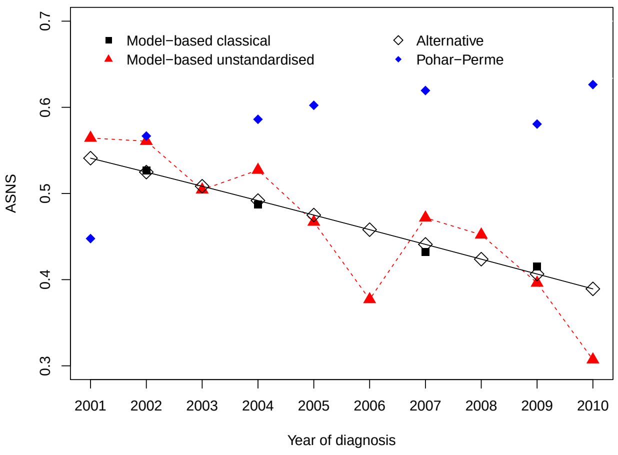

Figure 3: Vagina cancer 5-years age-standardised/unstandardised net survival by year of diagnosis.

Table 2: Results for scenarios A, B and C (samples size=200)

Modelling	Method	Model	Bias		Perc. Bias		Coverage		RMSE	
Strategy		convergency	2001-2010	2001	2001-2010	2001	2001-2010	2001	2001-2010	2001
					Scenario A					
FixedMod	*classical*	100.0	-0.0010	-0.0036	-0.3071	-1.1079	95.6	94.0	0.0352	0.0397
	alternative		-0.0015	-0.0015	-0.4581	-0.4581	95.5	95.5	0.0359	0.0359
BestMod	*classical*	100.0	-0.0046	-0.0077	-1.4252	-2.4128	95.5	91.9	0.0350	0.0388
	alternative		-0.0045	-0.0045	-1.4118	-1.4118	95.5	95.3	0.0350	0.0350
PP	-	-	-0.0022	-0.0149	-0.6933	-4.6481	94.9	93.9	0.0365	0.1101
					Scenario B					
FixedMod	*classical*	98.6	-0.0011	0.0002	-0.1256	0.0297	89.0	90.6	0.0440	0.0769
	alternative		0.0005	-0.0022	0.0622	-0.2713	90.1	91.4	0.0446	0.0754
BestMod	*classical*	100.0	0.0133	0.0095	1.5692	1.1646	89.6	89.5	0.0472	0.0858
	alternative		0.0115	0.0004	1.3471	0.0505	91.5	88.9	0.0466	0.0905
PP	-	-	-0.0053	-0.0192	-0.6255	-2.3615	93.1	88.7	0.0463	0.1010
					Scenario C					
FixedMod	*classical*	100.0	0.0024	-0.0224	0.3893	-3.8146	94.2	93.2	0.0411	0.0861
	alternative		0.0019	-0.0214	0.3082	-3.6271	95.2	94.0	0.0420	0.0828
BestMod	*classical*	100.0	0.0051	-0.0064	0.8059	-1.0880	94.9	93.6	0.0393	0.0868
	alternative		0.0040	-0.0075	0.6333	-1.2695	95.0	93.8	0.0397	0.0838
PP	-	-	-0.0009	-0.0234	-0.1367	-3.9887	95.5	93.6	0.0423	0.1286

Estimation of age-standardised net survival with sparse data: taking advantage of regression models.

Supplementary Material S1: Estimation of age-standardised net survival variance

Age-standardised net survival ($ASNS$) is calculated as a weighted average of age group specific net survivals ($NS_j(t)$):

$$ASNS(t) = \sum_{j=1}^{k} w_j \cdot NS_j(t)$$

where k is the number of age groups considered and w_j are the respective weights ($\sum w_j = 1$).

The age group specific net survival is obtained by averaging the individual NS predictions ($S_i(t)$):

$$NS_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} S_i(t)$$

where n_j is the number of patients in age group j.

The individual net survivals are obtained by integrating the excess hazard function:

$$S_i(t) = exp\left\{ - \int_0^t \lambda_{E_i}(u)du \right\}$$

We assumed a general flexible parametric model for the excess hazard function with a B-spline function for the log-baseline hazard, J variables with a time fixed-effect ($\mathbf{X} = \{X_1, X_2, \cdots, X_J\}$) and L variables with a time-dependent effect ($\mathbf{Z} = \{Z_1, Z_2, \cdots, Z_L\}$):

$$\begin{aligned}
\lambda_{E_i}(t, \beta) &= exp\left(\underbrace{\sum_{d=1}^{D} \beta_d \cdot Bspline_d(t) +}_{\text{log-baseline hazard}} \underbrace{\sum_{j=1}^{J} \beta_j \cdot X_j}_{\text{time-fixed effect}} \underbrace{+\sum_{l=1}^{L} Z_l \cdot \left(\sum_{d=1}^{D} \beta_{ld} \cdot Bspline_d(t) \right)}_{\text{time-dependent effect}} \right) \\
&= exp\left(\sum_{d=1}^{D} \left(\beta_d + \sum_{l=1}^{L} \beta_{ld} \cdot Z_l \right) \cdot Bspline_d(t) + \sum_{j=1}^{J} \beta_j \cdot X_j \right)
\end{aligned}$$

The vectors **X** and **Z** can share common covariates or interactions between covariates.

The age-standardised net survival is thus given by:

$$\begin{aligned} ASNS(t,\beta) &= \sum_{j=1}^{k} w_j \cdot NS_j(t,\beta) \\ &= \sum_{j=1}^{k} \left(w_j \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} S_i(t,\beta) \right) \\ &= \sum_{j=1}^{k} \left(\frac{w_j}{n_j} \cdot \sum_{i=1}^{n_j} S_i(t,\beta) \right) \end{aligned}$$

The variance of the age-standardised net survival can be obtained using the delta method[1]:

$$VAR\left(ASNS(t,\hat{\beta})\right) = \left[\frac{\partial ASNS(t,\beta)}{\partial \boldsymbol{\beta}}\right]_{\beta=\hat{\beta}} \times \left[VAR(\hat{\beta})\right] \times \left[\frac{\partial ASNS(t,\beta)}{\partial \boldsymbol{\beta}}\right]^{T}_{\beta=\hat{\beta}}$$

where

$$\left[\frac{\partial ASNS(t,\beta)}{\partial \boldsymbol{\beta}}\right] = \left[\frac{\partial}{\partial \beta} \sum_{j=1}^{k} \left(\frac{w_j}{n_j} \cdot \sum_{i=1}^{n_j} S_i(t,\beta) \right)\right] = \left[\sum_{j=1}^{k} \left(\frac{w_j}{n_j} \cdot \sum_{i=1}^{n_j} \frac{\partial S_i(t,\beta)}{\partial \beta} \right)\right]$$

So, to estimate the variance of the age-standardised net survival at a specific time t_1, considering a model with a total of $P(= J + L + D)$ parameters, it is necessary to solve the following product of matrices:

$$VAR(ASNS(t,\beta)) = \quad \left[\frac{\partial(ASNS(t,\beta))}{\partial\beta_1}(t_1) \quad \frac{\partial(ASNS(t,\beta))}{\partial\beta_2}(t_1) \quad \dots \quad \frac{\partial(ASNS(t,\beta))}{\partial\beta_P}(t_1)\right] \times$$

[1] *Delta method*

$$VAR(g(X)) = \left(\frac{\partial g(X)}{\partial X}\right)^2 \cdot VAR(X)$$

If dimension of X is greater than 1:

$$VAR(g(\boldsymbol{X})) = \left[\frac{\partial g(X)}{\partial X}\right] \times [VAR(\boldsymbol{X})] \times \left[\frac{\partial g(X)}{\partial X}\right]^T$$

$$\times \begin{bmatrix} VAR(\beta_1) & COV(\beta_1,\beta_2) & \ldots & COV(\beta_1,\beta_P) \\ COV(\beta_2,\beta_1) & VAR(\beta_2) & \ldots & COV(\beta_2,\beta_P) \\ \vdots & \vdots & \ddots & \vdots \\ COV(\beta_P,\beta_1) & COV(\beta_P,\beta_2) & \ldots & VAR(\beta_P) \end{bmatrix} \times$$

$$\times \left[\tfrac{\partial(ASNS(t,\beta))}{\partial\beta_1}(t_1) \quad \tfrac{\partial(ASNS(t,\beta))}{\partial\beta_2}(t_1) \quad \ldots \quad \tfrac{\partial(ASNS(t,\beta))}{\partial\beta_P}(t_1) \right]^T$$

The partial derivatives of the survival function with respect to each regression parameter of the time-fixed effects are given by:

$$\begin{aligned}
\frac{\partial S_i(t,\beta)}{\partial\beta_k} &= \\
&= \frac{\partial}{\partial\beta_k}\left(exp\left\{ -\int_0^t \lambda_{E_i}(u,\beta)du \right\}\right) \\
&= \frac{\partial}{\partial\beta_k}\left(exp\left\{ -\int_0^t exp\left(\sum_{d=1}^{D}\left(\beta_d + \sum_{l=1}^{L}\beta_{ld}\cdot z_l\right)\cdot Bspline_d(u) + \sum_{j=1}^{J}\beta_j \cdot X_j \right) du \right\}\right) \\
&= S_i(t,\beta)\cdot \frac{\partial}{\partial\beta_k}\left\{ -\int_0^t exp\left(\sum_{d=1}^{D}\left(\beta_d + \sum_{l=1}^{L}\beta_{ld}\cdot z_l\right)\cdot Bspline_d(u) + \sum_{j=1}^{J}\beta_j \cdot X_j \right) du \right\} \\
&= -S_i(t,\beta)\cdot \int_0^t \frac{\partial}{\partial\beta_k}\left\{ exp\left(\sum_{d=1}^{D}\left(\beta_d + \sum_{l=1}^{L}\beta_{ld}\cdot z_l\right)\cdot Bspline_d(u) + \sum_{j=1}^{J}\beta_j \cdot X_j \right) du \right\} \\
&= -S_i(t,\beta)\cdot \int_0^t X_k \cdot \lambda_{E_i}(u,\beta)du
\end{aligned}$$

With respect to each regression parameter of the baseline hazard, the derivatives are given by:

$$\begin{aligned}
&\frac{\partial S_i(t,\beta)}{\partial \beta_{dk}} = \\
&\qquad = \frac{\partial}{\partial \beta_{dk}} \left(exp\left\{ - \int_0^t \lambda_{E_i}(u,\beta) du \right\} \right) \\
&\qquad = \frac{\partial}{\partial \beta_{dk}} \left(exp\left\{ - \int_0^t exp\left(\sum_{d=1}^{D} \left(\beta_d + \sum_{l=1}^{L} \beta_{ld} \cdot Z_l \right) \cdot Bspline_d(u) + \sum_{j=1}^{J} \beta_j \cdot X_j \right) du \right\} \right) \\
&\qquad = -S_i(t,\beta) \cdot \int_0^t \frac{\partial}{\partial \beta_{dk}} \left\{ exp\left(\sum_{d=1}^{D} \left(\beta_d + \sum_{l=1}^{L} \beta_{ld} \cdot Z_l \right) \cdot Bspline_d(u) + \sum_{j=1}^{J} \beta_j \cdot X_j \right) du \right\} \\
&\qquad = -S_i(t,\beta) \cdot \int_0^t Bspline_{dk}(u) \cdot \lambda_{E_i}(u,\beta) du
\end{aligned}$$

And with respect to each regression parameter of the time-dependent effect are given by:

$$\begin{aligned}
&\frac{\partial S_i(t,\beta)}{\partial \beta_{lk}} = \\
&\qquad = \frac{\partial}{\partial \beta_{lk}} \left(exp\left\{ - \int_0^t \lambda_{E_i}(u,\beta) du \right\} \right) \\
&\qquad = \frac{\partial}{\partial \beta_{lk}} \left(exp\left\{ - \int_0^t exp\left(\sum_{d=1}^{D} \left(\beta_d + \sum_{l=1}^{L} \beta_{ld} \cdot Z_l \right) \cdot Bspline_d(u) + \sum_{j=1}^{J} \beta_j \cdot X_j \right) du \right\} \right) \\
&\qquad = -S_i(t,\beta) \cdot \int_0^t \frac{\partial}{\partial \beta_{dk}} \left\{ exp\left(\sum_{d=1}^{D} \left(\beta_d + \sum_{l=1}^{L} \beta_{ld} \cdot Z_l \right) \cdot Bspline_d(u) + \sum_{j=1}^{J} \beta_j \cdot X_j \right) du \right\} \\
&\qquad = -S_i(t,\beta) \cdot \int_0^t Bspline_{lk}(u) \cdot Z_{lk} \cdot \lambda_{E_i}(u,\beta) du
\end{aligned}$$

Using again the delta method, the variance of the $log(-log(ASNS))$ is given by:

$$\begin{aligned}
&VAR\Big(log(-log(ASNS(t,\beta))) \Big) = \\
&\left[\frac{\partial}{\partial \beta} \Big(log(-log(ASNS(t,\beta))) \Big) \right]_{\beta=\hat{\beta}} \times \Big[VAR(\hat{\beta}) \Big] \times \left[\frac{\partial}{\partial \beta} \Big(log(-log(ASNS(t,\beta))) \Big) \right]_{\beta=\hat{\beta}}^{T}
\end{aligned}$$

where

$$\left[\frac{\partial}{\partial\beta}\Big(log(-log(ASNS(t,\beta)))\Big)\right] = \left[\frac{\frac{\frac{\partial}{\partial\beta}\Big(ASNS(t,\beta)\Big)}{ASNS(t,\beta)}}{log(ASNS(t,\beta))}\right] = \left[\frac{\sum_{j=1}^{k}\left(\frac{w_j}{n_j}\cdot\sum_{i=1}^{n_j}\frac{\partial S_i(t,\beta)}{\partial\beta}\right)}{ASNS(t,\beta)\cdot log(ASNS(t,\beta))}\right]$$

So,

$$VAR\Big(log(-log(ASNS(t,\beta)))\Big) = \frac{VAR(ASNS(t,\beta))}{ASNS(t,\beta)^2\cdot log(ASNS(t,\beta))^2}$$

Assuming normality of $log(-log(ASNS))$, the confidence interval for the age-standardised net survival ($ASNS$) is given by:

$$\begin{aligned} CI_{100(1-\alpha)\%} : & exp\left(-exp\Big(log(-log(ASNS)) \pm Z_{\alpha/2}\cdot\hat{\sigma}_{log(-log(ASNS))}\Big)\right) \\ & exp\left(-exp\Big(log(-log(ASNS)) \pm Z_{\alpha/2}\cdot\frac{\hat{\sigma}_{ASNS}}{ASNS\cdot log(ASNS)}\Big)\right) \end{aligned}$$

Estimation of age-standardised net survival with sparse data: taking advantage of regression models.

Supplementary Material S2: Model selection algorithm

In the second modelling strategy used in the simulation study, a backward model-building algorithm based on the one proposed by Wynant and Abrahamowicz (Wynant and Abrahamowicz (2014)) was used.

In scenario A no effect of year of diagnosis was considered. The list of potential models from which the final model was selected was:

MA1 $log[\lambda_E(a,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + a \cdot f(t)$

MA2 $log[\lambda_E(a,t)] = log(\lambda_0(t)) + a + a \cdot f(t)$

MA3 $log[\lambda_E(a,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3$

MA4 $log[\lambda_E(a,t)] = log(\lambda_0(t)) + a$

where a represents age (centred and scaled) and k is a fixed knot.

Model-building strategy:

i Test linearity of age assuming time-dependent effect and test time-dependent effect assuming non-linearity of age by comparing MA1/MA2 (p-value → p_{12}) and MA1/MA3 (p-value → p_{13}), using likelihood ratio tests:

ii If $p_{12} < \alpha$ and $p_{13} < \alpha$, choose MA1.

iii If $p_{12} > p_{13}$ and $p_{12} \geq \alpha$, test time-dependent effect assuming linearity of age by comparing MA2/MA4 (p-value → p_{24}):

- if $p_{24} \geq \alpha$, choose MA4;
- if $p_{24} < \alpha$, choose MA2.

iv If $p_{13} > p_{12}$ and $p_{13} \geq \alpha$, test linearity of age assuming proportional hazards by comparing MA3/MA4 (p-value → p_{34}):

- if $p_{34} \geq \alpha$, choose MA4;
- if $p_{34} < \alpha$, choose MA3.

In scenario B and C, a linear effect of year of diagnosis was always considered. Time-dependent effect of year and interaction between age and year were considered in the list of potential models from which the final model was selected:

M1 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y + a*y + (a+y) \cdot f(t)$

M2 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y + (a+y) \cdot f(t)$

M3 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y + a*y + (a) \cdot f(t)$

M4 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y + a*y + (y) \cdot f(t)$

M5 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y + a*y$

M6 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y + a*y + (a+y) \cdot f(t)$

M7 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y + a*y + (a) \cdot f(t)$

M8 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y + a*y + (y) \cdot f(t)$

M9 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y + a*y$

M10 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y + (a) \cdot f(t)$

M11 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y + (y) \cdot f(t)$

M12 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + a^2 + a^3 + I(a > k) \cdot (a-k)^3 + y$

M13 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y + (a+y) \cdot f(t)$

M14 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y + (a) \cdot f(t)$

M15 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y + (y) \cdot f(t)$

M16 $log[\lambda_E(a,y,t)] = log(\lambda_0(t)) + a + y$

The model-building strategy used in these scenarios was as follows:

i Test interaction between age and year of diagnosis by comparing models M1/M2 (p_{12}):

- if $p_{12} < \alpha$, choose models with interaction.
- if $p_{12} \geq \alpha$, choose models without interaction.

ii Test time-dependent effect of year, assuming time-dependent effect of age (M1/M3 or M2/M10), test time-dependent effect of age, assuming time-dependent effect of year (M1/M4 or M2/M11) and test non-linearity of age assuming time-dependent effect of age (M1/M6 or M2/M13);

- If all effects are significant, choose model M1 (interaction) or model M2 (no interaction);
- If any $p - value \geq \alpha$, remove least significant effect;

iii Continue removing the least significant effects of the successively simpler models until finding a model with only significant effects;

iv The simplest possible model assumes linear effect of age and year of diagnosis and proportional hazards of both variables.

References

Wynant, W. and Abrahamowicz, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in Medicine*, **33**(19), 3318–3337. ISSN 10970258. doi: 10.1002/sim.6178.

Estimation of age-standardised net survival with sparse data: taking advantage of regression models.

Supplementary Material S3: Results for samples $n = 2000$

Table 1: Results for scenarios A, B and C (samples size=2000)

Modelling	Method	Model	Bias		Perc. Bias		Coverage		RMSE	
Strategy		convergency	2001-2010	2001	2001-2010	2001	2001-2010	2001	2001-2010	2001
			Scenario A							
FixedMod	*classical*	100.0	-0.0004	-0.0004	-0.1229	-0.1351	94.3	94.5	0.0114	0.0116
	alternative		-0.0005	-0.0005	-0.1439	-0.1439	93.1	93.1	0.0116	0.0116
BestMod	*classical*	100.0	-0.0015	-0.0015	-0.4541	-0.4687	93.4	93.9	0.0117	0.0119
	alternative		-0.0015	-0.0015	-0.4604	-0.4604	93.1	93.0	0.0118	0.0118
PP	-	-	-0.0006	-0.0087	-0.1964	-2.7304	94.8	93.6	0.0118	0.0368
			Scenario B							
FixedMod	*classical*	100.0	0.0005	-0.0003	0.0536	-0.0376	94.5	94.2	0.0141	0.0234
	alternative		0.0005	0.0000	0.0586	0.0009	94.7	94.0	0.0139	0.0228
BestMod	*classical*	100.0	0.0028	0.0001	0.3298	0.0181	89.1	91.1	0.0158	0.0278
	alternative		0.0022	-0.0022	0.2621	-0.2662	91.8	91.7	0.0149	0.0271
PP	-	-	-0.0008	-0.0162	-0.0983	-1.9836	94.7	91.7	0.0149	0.0493
			Scenario C							
FixedMod	*classical*	100.0	0.0000	-0.0014	0.0024	-0.2383	93.9	94.1	0.0132	0.0250
	alternative		-0.0001	-0.0017	-0.0159	-0.2839	93.9	94.4	0.0135	0.0250
BestMod	*classical*	100.0	0.0059	0.0052	0.9369	0.8919	90.9	93.3	0.0152	0.0273
	alternative		0.0047	0.0029	0.7481	0.5004	91.0	94.0	0.0152	0.0275
PP	-	-	-0.0014	-0.0007	-0.2257	-0.1242	94.5	95.2	0.0139	0.0430

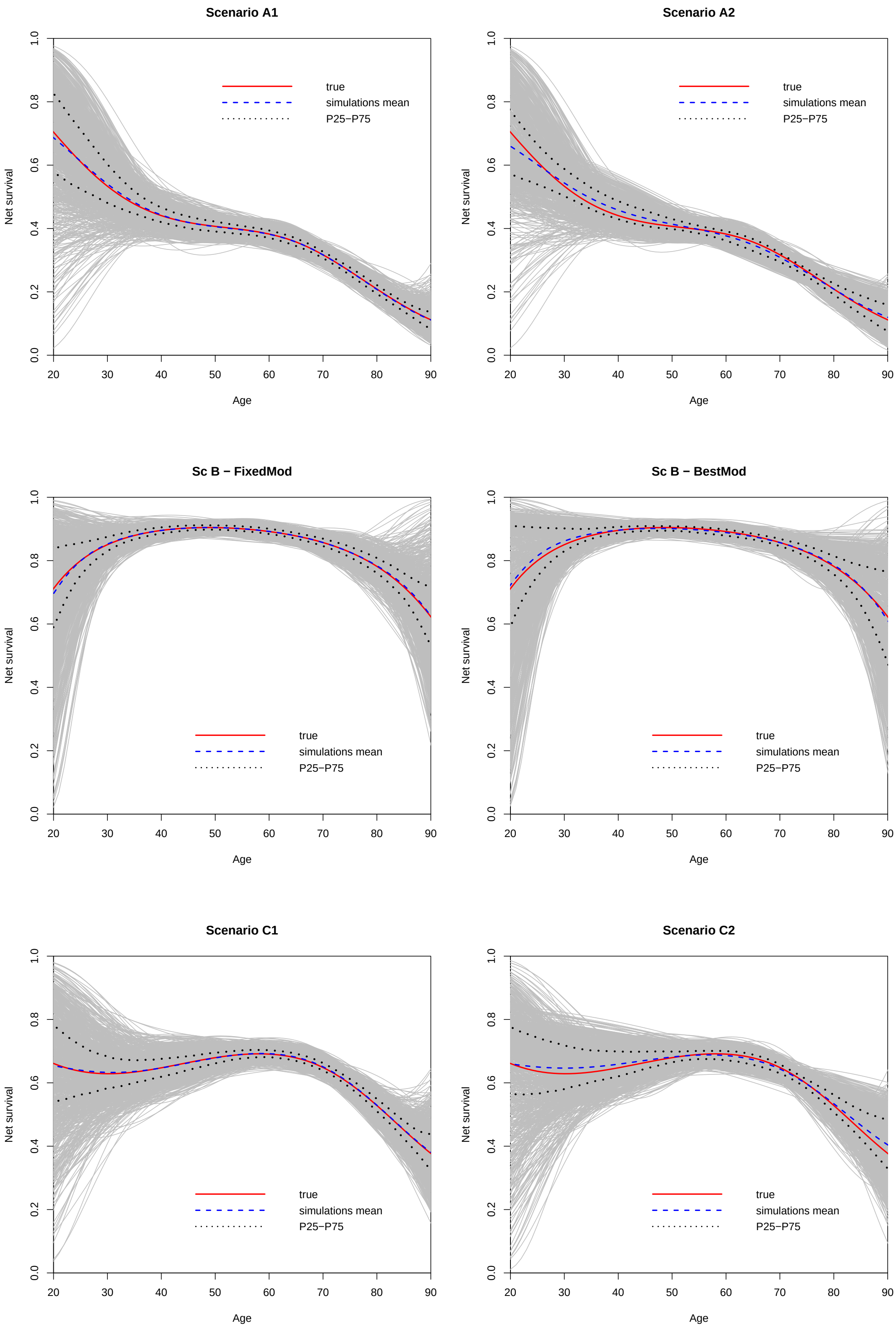

Figure 1: Predicted survival by age: a) Scenario A; b) Scenario B; c) Scenario C. Model fitted to samples of size $n = 2000$

3.2 Study II: Socioeconomic inequalities in survival from cancer

No inequalities in survival from colorectal cancer by education and socioeconomic deprivation - a population-based study in the North Region of Portugal, 2000-2002
Luis Antunes, Denisa Mendonça, Maria José Bento, Bernard Rachet
BMC Cancer 2016 16:608

Survival from cancer has been shown to be frequently associated to socioeconomic factors. Although most of the published studies point to a more favourable prognosis to less deprived patients, there are also some studies that did not find any association. In this study, an evaluation of the association of socioeconomic status of cancer patients with their survival from the disease was for the first time in Portugal performed. The study focused on colorectal cancer patients diagnosed in the period 2000-2002 in the North region of Portugal.

Area-based variables were used for the attribution of socioeconomic condition. A single indicator (education) and a composite index (European Deprivation Index) were used as socioeconomic variables. The geographical unit considered was the section tract (*Secção estatística*). All the patient's addresses were geocoded and after matched with the relevant census area using a Geographical Information System.

Age-standardised net survival was estimated using the Pohar-Perme non-parametric estimator, by socioeconomic group, stratifying by sex. Excess hazard ratios were estimated using flexible parametric models, considering time-dependent effects of the socioeconomic variable.

In this study, general life tables were used since no deprivation-specific life tables were available at the time the study was performed. Subsequent sensitivity analysis to socioeconomic differences in background mortality was performed. Although some cancer survival inequalities were found for men when using general lifetimes, the sensitivity analysis showed that small deprivation gaps in background mortality cancelled out the inequalities observed in the survival from the disease.

Next, the resulting manuscript of this study is presented. This has been published in *BMC Cancer* in 2016.

BMC Cancer

RESEARCH ARTICLE Open Access

No inequalities in survival from colorectal cancer by education and socioeconomic deprivation - a population-based study in the North Region of Portugal, 2000-2002

Luís Antunes[1,2,3], Denisa Mendonça[4,5], Maria José Bento[1,2,6] and Bernard Rachet[7*]

Abstract

Background: Association between cancer survival and socioeconomic status has been reported in various countries but it has never been studied in Portugal. We aimed here to study the role of education and socioeconomic deprivation level on survival from colorectal cancer in the North Region of Portugal using a population-based cancer registry dataset.

Methods: We analysed a cohort of patients aged 15–84 years, diagnosed with a colorectal cancer in the North Region of Portugal between 2000 and 2002. Education and socioeconomic deprivation level was assigned to each patient based on their area of residence. We measured socioeconomic deprivation using the recently developed European Deprivation Index. Net survival was estimated using Pohar-Perme estimator and age-adjusted excess hazard ratios were estimated using parametric flexible models. Since no deprivation-specific life tables were available, we performed a sensitivity analysis to test the robustness of the results to life tables adjusted for education and socioeconomic deprivation level.

Results: A total of 4,105 cases were included in the analysis. In male patients (56.3 %), a pattern of worse 5- and 10-year net survival in the less educated (survival gap between extreme education groups: -7 % and -10 % at 5 and 10 years, respectively) and more deprived groups (survival gap between extreme EDI groups: -5 % both at 5 and 10 years) was observed when using general life tables. No such clear pattern was found among female patients. In both sexes, when likely differences in background mortality by education or deprivation were accounted for in the sensitivity analysis, any differences in net survival between education or deprivation groups vanished.

Conclusions: Our study shows that observed differences in survival by education and EDI level are most likely attributable to inequalities in background survival. Also, it confirms the importance of using the relevant life tables and of performing sensitivity analysis when evaluating socioeconomic inequalities in cancer survival. Comparison studies of different healthcare systems organization should be performed to better understand its influence on cancer survival inequalities.

Keywords: Net survival, Colorectal cancer, Education, Deprivation, Inequalities, Life tables

* Correspondence: bernard.rachet@lshtm.ac.uk
[7]Cancer Survival Group, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Full list of author information is available at the end of the article

Background

Colorectum is the second most common cancer site in the North Region of Portugal, only surpassed by prostate in men and breast in women [1]. Age-standardized incidence rates of both colon and rectal cancers have been recently rising in this region of Europe and are predicted to continue rising, at least until 2020 [2, 3]. Five-year survival from colorectal cancer (CRC) in Portugal was generally higher than in Eastern European countries, the UK, Denmark and Spain, and lower than in The Netherlands, France, Italy and the Nordic countries among others [4].

Association between survival from colon or rectal cancer and socioeconomic status (SES) has been repeatedly reported in various countries [5, 6]. Socioeconomic condition can be attributed to each patient using individual measures [7–9]. However, population-based cancer registries rarely collect individual data on socioeconomic factors. Alternatively, ecological (area-based) measures are used [10–12]. Although not reflecting the individual condition of each patient, ecological measures are informative enough to evaluate the association between SES and survival from cancer, as long as the population size of the areas considered is sufficiently small and homogeneous relatively to the SES measure [13]. The SES can be measured using single indicators (e.g., income, education) [9, 14] or composite indices (e.g., Townsend, Indices of Multiple Deprivation) [10, 11, 15]. Because the large number of different indicators found in the literature can hamper comparisons between studies, a new ecological socioeconomic deprivation index (European Deprivation Index – EDI) has been recently developed for several European countries (Portugal, Spain, France, Italy, England), based on the same methodology across all countries [16]. The index is derived from country-specific census variables that are most associated with the variables of the survey European Union-Statistics on Income and Living Conditions EU-SILC [17].

Independently of the SES measure, patients with a lower SES are generally found to present a worse survival compared to patients with a higher SES. Potential reported causes for SES inequalities in survival include variations in stage of disease at presentation, type of treatment delivered or patient characteristics [6, 18].

The National Health Service (SNS) functions in Portugal since 1979 and aims to provide the population with complete and high-quality care, independently of their social or economic condition. Cancer patients were totally exempt of paying moderating fees until the end of 2011. In an evaluation of the Portuguese situation regarding CRC, Pinto and colleagues suggested that one of the major problems in the management of the diagnosis and treatment of colorectal cancer patients were regional disparities in access to health [19]. However, to the best of our knowledge, socioeconomic inequalities in cancer survival in Portugal have not been assessed yet.

In the present study we aimed at evaluating the association between up-to-10-year survival from colorectal cancer and two indicators: the recently developed area-based socioeconomic indicator EDI and education level based on census information. We used population-based data from the North Region Cancer Registry of Portugal (RORENO).

Methods

Cancer registry

Cancer data were provided by RORENO, a population-based cancer registry established in 1988. The analyses were performed according to RORENO guidelines ensuring the anonymity of the information used. Its catchment area corresponds to the North Region of Portugal, with 3.2 million inhabitants (around 30 % of the national population). All incident cancer cases occurring in the area were recorded by the registry either directly from the main public hospitals through a web-based platform, or based on the hard copies of the medical reports for the private hospitals and pathology laboratories. Registration quality follows IARC rules [20].

Data

We considered for analysis all malignant, invasive tumours of the colon and rectum (ICD-10 [21] codes C18-20) diagnosed in adults resident in the North Region of Portugal in the period 2000 to 2002. For patients diagnosed with more than one tumour during the study period, only the first primary tumour contributed to the analysis. Follow-up of each patient was both active (by contacting the institutions where the patient was diagnosed and/or treated) and, when necessary, passive (by obtaining the vital status from the National Health Service database or the Civil Registration Offices). The end of follow-up was 31st December 2012, allowing over 10 years of potential follow-up for all patients. Because 10-year net survival is meaningless for very old patients, the study was restricted to patients aged 15 to 84 years.

Education and EDI level

No information on education or other SES indicator at individual level is systematically registered by cancer registries in Portugal. Education level and the socioeconomic deprivation index (EDI) were assigned to each patient based on their census area of residence at diagnosis. When not available, patient's address was completed using the National Health Service database. The residence of each patient was geocoded using a web-based service [22] and then confirmed manually. The coordinates of each patient's address were then matched with the relevant census area using a Geographical Information System (Arc GIS 10.2).

Education level was measured as the proportion of inhabitants in each census area aged 15 years or plus with at least 9 years of education (compulsory level of education in Portugal until 2009). This information was retrieved from the 2001 national census and the census area (in Portuguese: *secção estatística*) corresponds to the area of a census taker [23] (median population size: 665; range: 13 – 3123; number of sections: 4651). Education level was then categorized in five levels according to the quintiles of the regional distribution of all area-level education proportions. The distribution was weighted by the population size in each census area so that each level corresponds to 20 % of the total population (and not to 20 % of the number of sections). The first category corresponds to the census areas with the lowest proportion of residents with at least the compulsory level of education (proportion lower than 18.0 %) and the fifth category to areas with the highest proportion (proportion equal or higher than 48.9 %). The EDI was attributed to the census areas and categorized in five groups from q1 (the most deprived) to q5 (the least deprived).

Statistical analysis

Age distribution between groups was compared using Kruskall-Wallis or Mann–Whitney non parametric tests, as applicable. Survival time was considered as the time between diagnosis and death from any cause or end of study period, whichever occurred first. Up-to-10-year net survival was estimated using the Pohar-Perme non-parametric estimator [24]. Net survival is the survival that would be observed if cancer was the only possible cause of death and can be interpreted as the survival from the cancer. To this purpose, it accounts for the other causes of death or expected mortality. Within the relative survival setting, i.e., when the individual cause of death is not reliably known, the background or expected mortality is provided by life tables for the general population, here of the North Region of Portugal. The tables were built by the Cancer Survival Group (London School of Hygiene and Tropical Medicine) for the CONCORD-2 programme [25], using a multivariable flexible Poisson model [26]. The population and death counts to derive the life tables were provided by the national statistics office (Statistics Portugal). Life tables were stratified by sex, single year of age and calendar year.

Excess (i.e., cancer-related) hazards of death are also of interest. Univariable excess hazard models were used to test significance of potential prognostic variables (sex, age group, cancer site). Multivariable flexible parametric models [27] were used to estimate the hazard ratios of excess mortality for education and EDI levels, adjusted for potential confounders. Men and women were analysed separately. Education level and deprivation were kept in the model as categorical variables. Different models for the effect of age on the excess hazard were tested, considering age as categorical or continuous variable, with possible non-linear effect using restricted cubic splines. Time-dependent effects for age, education and EDI level were tested. The model with the lowest Akaike Information Criterion (AIC) was chosen.

All analyses were performed using STATA commands *stns* [28] and *stpm2* [29]. Results were considered statistically significant for p-value < 0.05.

Sensitivity analysis

Socioeconomic condition can affect the mortality of a cancer patient from both their cancer and other causes. Assessing socioeconomic inequalities in cancer survival should therefore account for socioeconomic differences in mortality from other causes (the expected or background mortality) [30]. Ignoring such differences leads to over-estimate the inequalities in cancer survival. Since no education-specific neither EDI-specific life tables are available in Portugal, we performed a sensitivity analysis to test the robustness of the results to the choice of the life tables. We built a series of hypothetical SES-specific life tables for Portugal according to various scenarios of inequalities in background mortality. Under the worst case scenario, we mimicked the wide gap in background mortality observed between socioeconomic categories in England, as illustrated by the English (2001) deprivation-specific life tables (http://csg.lshtm.ac.uk/). We further refer to this scenario as S5, and the scenario with no gap as S0. The worst case scenario (S5) corresponded to a difference in life expectancy between extreme groups of 7.7 years in men and of 4.1 years in women. In scenario S4, we reduced the difference in background mortality between SES groups by 20 %, obtaining a gap in life expectancy of 6.2 and 3.3 years in men and women, respectively. We continue reducing the gap in 20 % steps to produce the other life table sets (S3, S2, S1 with corresponding differences in life expectancy at 4.6, 3.1 and 1.5 years in men, and 2.5, 1.7 and 0.8 years in women) until the gap vanishes (S0). We then re-ran the survival analysis using each of these life tables.

Results

We identified 4,243 cases of colorectal cancer eligible for analysis over the period 2000-2002. After excluding cases with missing information on their vital status at the end of follow-up (n = 113) or on their residence address (n = 25), 4,105 cases (96.7 %) were included in the analysis (Table 1). More than half (56.3 %) were male. Distribution of age at diagnosis was similar in both sexes

Table 1 Description of the cases included in the analysis stratified by sex

Variable	Male		Female	
	n	%	*n*	%
All	2310	100	1795	100
Age group				
15–44	114	4.9	125	7.0
45–54	268	11.6	209	11.6
55–64	548	23.7	364	20.3
65–74	876	37.9	616	34.3
75–84	504	21.8	481	26.8
Education level				
Higher education	516	22.3	434	24.2
q4	543	23.5	422	23.5
q3	475	20.6	366	20.4
q2	400	17.3	328	18.3
Lower education	376	16.3	245	13.6
EDI				
Least deprived	377	16.3	289	16.1
q4	403	17.4	310	17.3
q3	459	19.9	334	18.6
q2	490	21.2	393	21.9
Most deprived	581	25.2	469	26.1
Cancer site				
Colon	1421	61.5	1206	67.2
Rectum	889	38.5	589	32.8

(median 68 years, interquartile range 59-74). Colon cancer patients represented nearly two thirds of the cases (64.0 %) and were slightly older than rectal cancer patients (median 68 versus 67 years, p-value = 0.002). The proportion of colorectal cancer patients increased towards the more educated groups. The distribution of patients by EDI level was in the opposite direction, with a higher proportion in the more deprived groups. Median age ranged from 67 to 68 years in the highest and least educated groups (p-value = 0.176), and from 66 to 68 years between the least and most EDI deprived groups (p-value = 0.056).

Net survival at 1, 5 and 10 years since diagnosis was 81.5 % (95%CI: 80.3–82.8), 57.5 % (95%CI: 55.7–59.3) and 51.6 % (95%CI: 49.4–53.8), respectively. No significant differences in net survival were found by sex (p-value = 0.460) or cancer site (p-value = 0.209). Net survival was significantly lower in elderly patients (aged 75-84 years) than in the youngest age group (p-value < 0.001) while no significant differences were found among all other age groups. This pattern was similar in both genders and for both cancer sites (data not shown).

For male patients, 1-year net survival estimated using general life tables was similar across education categories, ranging from 80 % to 83 % (Table 2). However, there was an education-related pattern for longer-term survival. The gap in 5- and 10-year survival widened (Fig. 1a), with differences between the two extreme education groups at 7 % and 10 %, respectively. The gradient in net survival by EDI category was not as clear as by education quintile (Fig. 1b). Nevertheless, male patients coming from the least deprived group presented

Table 2 Net survival by education and EDI level at 1, 5 and 10 years after diagnosis[a]

	Male						Female					
	1-year		5-years		10-years		1-year		5-years		10-years	
	%	95 % CI	%	95 % CI	%	95 % CI	%	95 % CI	%	95 % CI	%	95 % CI
Education level												
Higher education	81	77 – 85	61	56 – 66	56	49 – 62	82	78 – 85	59	53 – 64	57	50 – 63
q4	80	77 – 84	59	54 – 64	52	46 – 58	83	79 – 86	54	49 – 60	50	44 – 56
q3	82	78 – 86	56	50 – 61	47	40 – 54	83	79 – 87	57	51 – 63	55	48 – 61
q2	82	78 – 86	55	50 – 61	46	39 – 53	85	80 – 89	65	59 – 71	56	48 – 63
Lower education	83	79 – 87	54	48 – 60	46	39 – 53	73	68 – 79	51	44 – 58	46	38 – 54
EDI												
Least deprived	81	77 – 85	60	54 – 66	53	46 – 60	83	78 – 88	60	54 – 67	58	50 – 65
q4	80	76 – 84	58	52 – 64	57	49 – 64	79	74 – 83	60	53 - 66	53	46 – 60
q3	82	78 – 85	59	54 – 65	48	41 – 54	81	77 – 86	56	50 – 62	49	42 – 56
q2	80	77 – 84	56	51 – 62	46	40 – 53	84	80 – 88	56	50 – 61	52	45 – 58
Most deprived	84	80 – 87	55	50 – 60	48	41 – 54	80	76 – 84	57	52 – 62	54	48 – 60

[a] Net survival estimated using general life tables

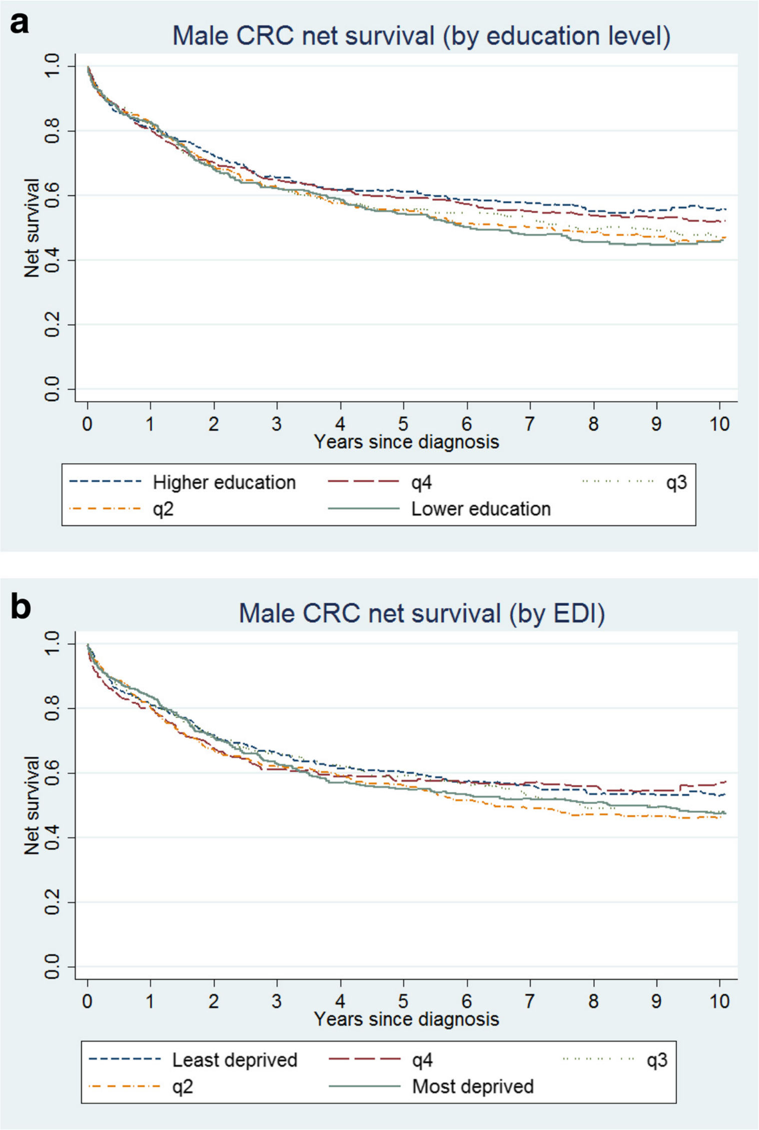

Fig. 1 Net survival for male patients: **a** by group of education level and **b** by EDI group (general Life Tables)

at 5 and 10 years a better net survival than patients coming from the most deprived groups.

By contrast, the pattern in survival across the five education levels was not gradual among women (Fig. 2a). Female patients coming from areas with the lowest education level presented always the lowest net survival over time. However, net survival hardly differed between the other education groups. Female net survival was also very similar between EDI groups, and not even the most deprived group detached from the remaining (Fig. 2b). Age-standardization of net survival estimates did not modify the survival pattern between education and EDI groups (Additional file 1: Table S1).

Adjusted excess hazard ratios (EHR) were computed from flexible parametric models with time-dependent effects for age and education and for age and EDI. We first used general life tables (i.e., not SES-specific). For male patients, the model confirmed the trend in increasing age-adjusted excess hazard across the education groups, more marked at longer term (Table 3). The excess hazard of death became significantly higher in the lowest educated group than in the highest educated (reference) group at 5 years (EHR = 1.40; 95 % CI: 1.06–1.84) and at 10 years (EHR = 1.51; 95 % CI: 1.08–2.11). For female patients, although the excess hazard in the lowest educated group was higher than the reference group, no statistically significant differences were found at 5 and 10 years since diagnosis (Table 3).

For male patients, the age-adjusted excess hazards for the more deprived groups were almost always higher

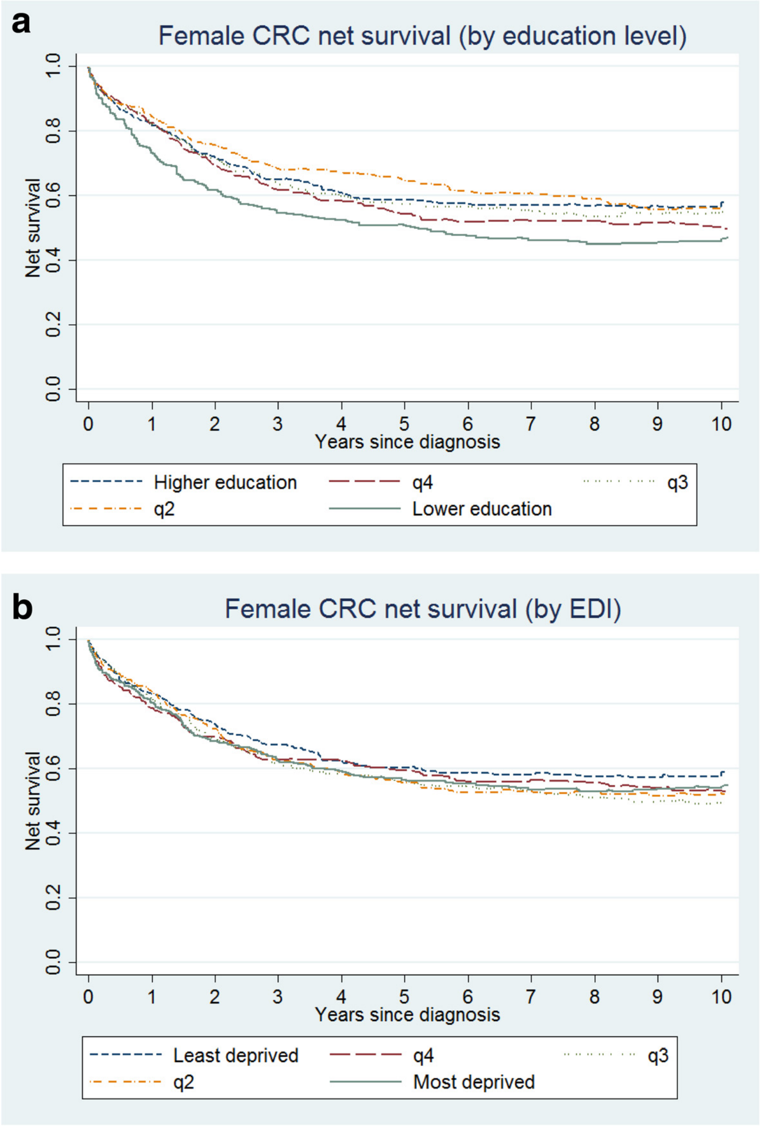

Fig. 2 Net survival for female patients: **a** by education level and **b** by EDI group (general Life Tables)

than the one observed for the reference group (least deprived). However, the EDI-related pattern of changes in excess hazard ratios was not as clear as with education. Again, no clear association between EDI and excess hazard was found among women.

To evaluate the sensitivity of the results, the excess hazard ratios by education and EDI level were re-estimated using different sets of life tables.

Overall, among men, the effect of education level variable was no longer significant in the excess hazard model as soon as fairly small inequalities in background mortality were considered (scenario S2). Figure 3 presents the excess hazard ratios at 5 (Fig. 3a) and 10 (Fig. 3b) years since diagnosis for the lowest education group, compared to the highest education group. Excess hazard at 5 and 10 years remained significantly different between the two extreme education groups only for narrow disparities in background mortality of the general population (scenarios S0 and S1). The excess hazard at 10 years of the least educated group was 51 % higher than the excess hazard of the group with highest education when using general life tables (S0). This difference reduced to 11 % when considering the English gap in background mortality (S5). For the EDI (Fig. 4a), the excess hazard ratio at 5 years between the most deprived group and the least deprived one reduced from 1.25 (S0) to less than one (S5). A similar behaviour was observed at 10 years (Fig. 4b).

Table 3 Excess Hazard Ratio estimates (and 95 % Confidence Intervals) by education level and EDI (adjusted for age)[a]

	Male						Female					
	1-year		5-years		10-years		1-year		5-years		10-years	
	HR	95 % CI	HR	95 % CI	HR	95 % CI	EHR	95 % CI	EHR	95 % CI	EHR	95 % CI
Education level												
Higher education	1		1		1		1		1		1	
q4	1.10	0.88 – 1.36	1.16	0.89 – 1.50	1.18	0.86 – 1.62	1.04	0.83 – 1.32	1.21	0.92 – 1.59	1.29	0.90 – 1.83
q3	1.15	0.92 – 1.43	1.27	0.97 – 1.65	1.32	0.95 – 1.82	1.02	0.81 – 1.30	1.05	0.78 – 1.41	1.06	0.73 – 1.55
q2	1.13	0.90 – 1.42	1.27	0.97 – 1.67	1.34	0.96 – 1.87	0.84	0.65 – 1.09	0.88	0.64 – 1.21	0.90	0.60 – 1.35
Lower education	1.16	0.92 – 1.46	1.40	1.06 – 1.84	1.51	1.08 – 2.11	1.33	1.03 – 1.71	1.27	0.93 – 1.75	1.25	0.83 – 1.87
EDI												
Least deprived	1		1		1		1		1		1	
q4	1.20	0.93 – 1.53	0.93	0.68 – 1.26	0.84	0.58 – 1.22	1.15	0.87 – 1.52	1.05	0.75 – 1.48	1.01	0.65 – 1.55
q3	1.04	0.81 – 1.33	1.07	0.80 – 1.43	1.08	0.76 – 1.55	1.10	0.83 – 1.44	1.26	0.91 – 1.74	1.33	0.88 – 2.01
q2	1.19	0.94 – 1.51	1.30	0.98 – 1.71	1.34	0.95 – 1.88	1.06	0.81 – 1.39	1.14	0.83 – 1.57	1.18	0.79 – 1.76
Most deprived	1.14	0.90 – 1.43	1.25	0.96 – 1.64	1.30	0.93 – 1.82	1.15	0.89 – 1.48	1.02	0.75 – 1.40	0.97	0.65 – 1.45

[a]Excess hazard ratios estimated using general life tables

Among women, as expected, the initial lack of inequalities observed with the general life tables remained for all scenarios (Additional file 2: Figures S1, S2).

Discussion

When the expected (or background) mortality of the cancer patients was provided by general life tables, net survival from colorectal cancer tended to decrease with decreasing education level in men. These inequalities however occurred only for long-term survival, i.e., at 5 and 10 year since diagnosis. No clear gradient was observed for women, in spite of a general worse survival in the less educated group.

Inequalities in survival were in general smaller by EDI level than by education. This was true for both genders.

General life tables assume that the patients have the same (age-, sex- and calendar year-specific) expected mortality, regardless their education or EDI level, which is unlikely. It may result in an overestimation of the survival gap [30], in particular as time since diagnosis is increasing, as illustrated by our results. In the absence of education-specific or EDI-specific life tables in Portugal, we performed a sensitivity analysis, using hypothetical life tables adjusted for the respective SES measure. This analysis revealed that differences in expected mortality reduced considerably the observed inequalities in net survival. Fairly small education-related differences in expected mortality (scenario S2 – Fig. 3a) were sufficient to cancel the inequalities in net survival between education groups initially observed (S0). Scenario S2 corresponds to a difference in life expectancy as small as 3.1 years between the most educated and least educated categories in the general population, a difference which is likely to underestimate the real disparities in background mortality between socioeconomic or education groups in Portugal (i.e., still to overestimate the cancer survival gap). The gap in life expectancy in that scenario is for example smaller than the difference (3.6 years) observed between the North Region and the Portuguese islands (Madeira and Azores) [31], where the lowest life expectancy at birth in Portugal is observed. Disparities in background mortality are plausible since there is also strong evidence of worse health status in more deprived classes. Higher prevalence of cardiovascular disease, stroke, ischemic heart disease, hypertension, diabetes, obesity and low physical inactivity has been associated with lower socioeconomic status in Portugal [32]. In the Metropolitan Area of Porto, increased early mortality rates have been shown in more deprived parishes [33].

Although the general conclusions were similar, results obtained with education and EDI differed. The analysis of the area typology reveals that education level seems to be more related to a rural/urban distinction than EDI. While about 40 % of the patients coming from the least educated areas live in rural areas, only 13 % of the patients living in the more deprived areas correspond to rural zones. Since the major treatment centres are in urban areas, this suggests that the least educated patients have a worse accessibility to treatment centres. This is in accordance with Pinto et al. [19] that identified regional disparities in access to health care facilities as one of the major problems in the management of diagnosis and treatment of colorectal cancer patients.

Differential participation rate in screening programmes by socioeconomic condition is a source of inequalities in survival. In the region considered in this study however, no organized CRC screening programme existed during

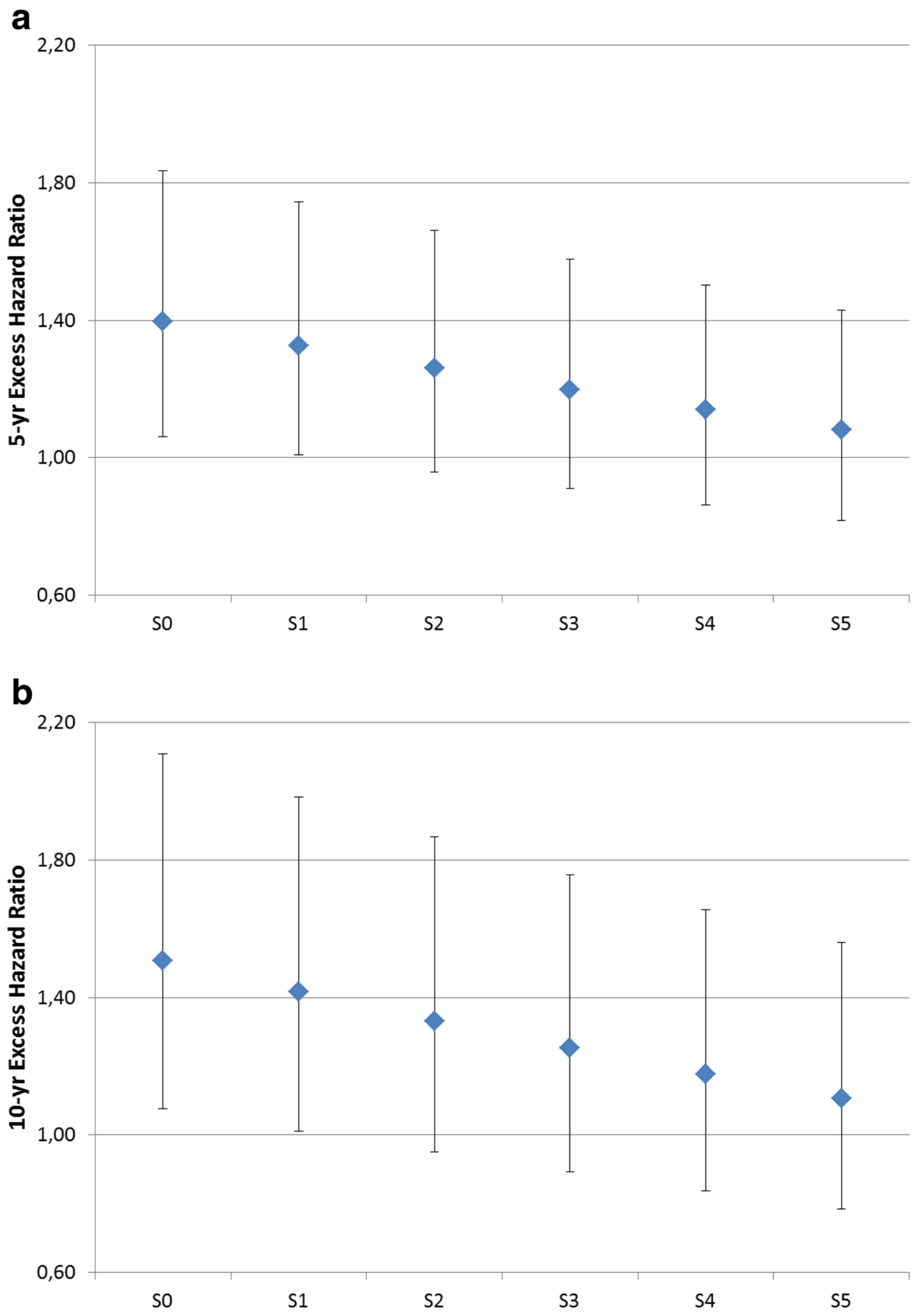

Fig. 3 Sensitivity analysis – Excess Hazard Ratios for the least educated group (compared with most educated group) at **a** 5 years and **b** 10 years since diagnosis (male patients)

the period of diagnosis analysed, neither is yet implemented at the present. In Portugal, an official pilot CRC screening programme was initiated in 2009 in the centre region. In 2014, CRC screening programmes covered only 3.7 % of the Portuguese population [34]. Participation in opportunistic screening remained also low: a questionnaire study performed in Porto municipality in 2009 showed that about two thirds of the inquired (mean age 60 years-old) had never performed any type of CRC screening exam [35]. This study found no association between the knowledge of CRC risk factors and education level.

The association between CRC and socioeconomic factors has been evaluated in different countries with different health care systems [5, 6]. Some methodological differences in published studies can be pointed out. First, socioeconomic condition is defined either at individual level [7–9] or using an ecological measure [11, 12, 14].

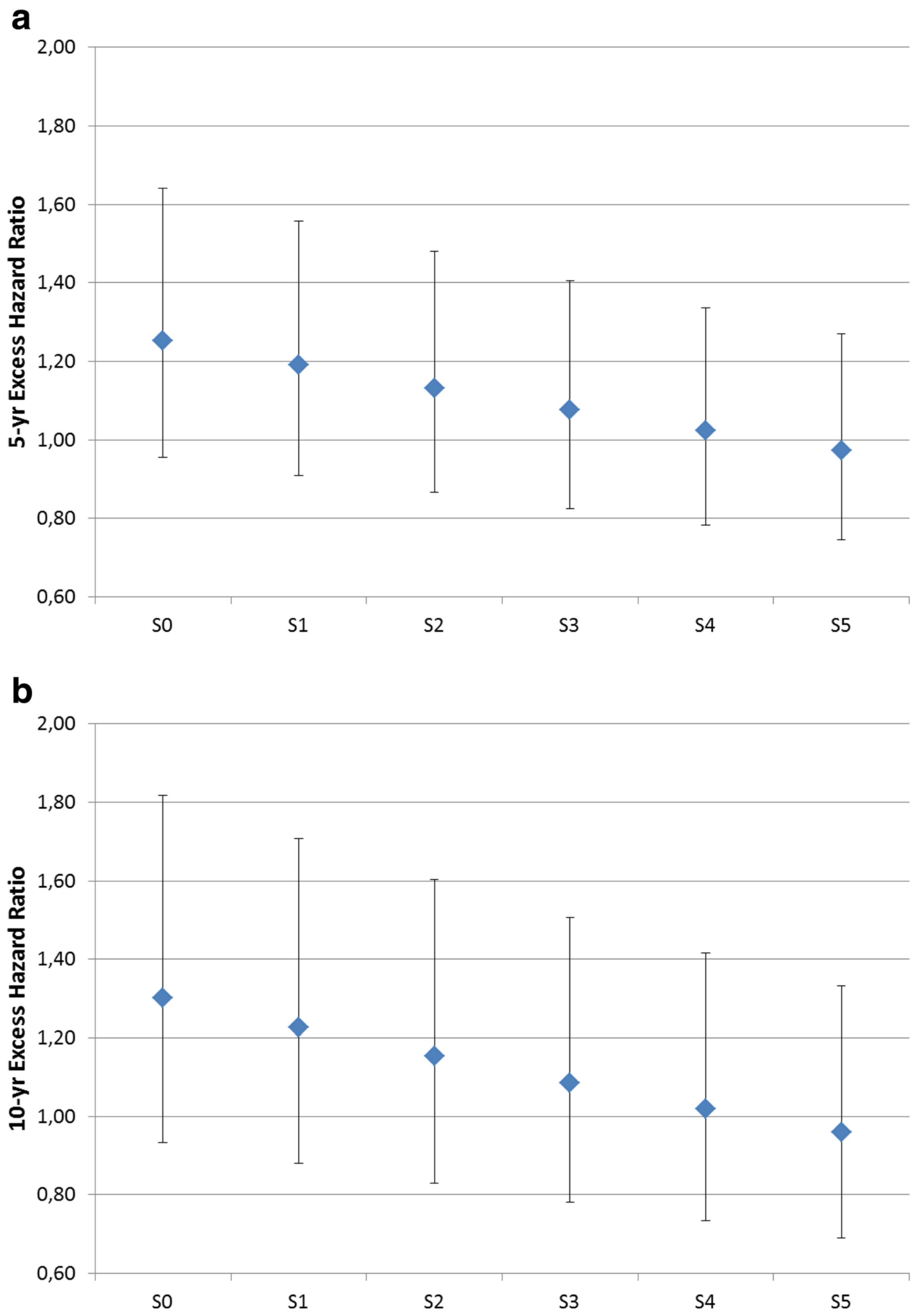

Fig. 4 Sensitivity analysis – Excess Hazard Ratios for the most deprived group (compared with least deprived group) at **a** 5 years and **b** 10 years since diagnosis (male patients)

Second, the metric to measure socioeconomic condition varies. Third, the outcome used is not homogeneous. Overall [36–38], cancer-specific [39, 40] or relative survival [7, 9, 41] have been used as outcome measures. Beyond these differences, most studies found an association between socioeconomic condition and survival from colorectal cancer.

A Danish study found a lower relative survival at 1 and 5 years for colon and rectum cancer patients with basic or high school, relatively to patients with vocational or higher education, for both genders [7]. Improved survival for more highly educated men was observed in Sweden for both colon and rectum cancers, compared with men with less than 9 years of completed education, while for women this difference was observed only for colon cancer [8]. Another study in Sweden found also a clear pattern of better survival for more highly educated groups [9]. Socioeconomic inequalities in colon and/or rectum cancer survival have also been found in England [11] and Japan [12]. Gorey and coworkers evaluated the association between income and colon cancer survival in San Francisco (US) and Toronto (Canada) [36]. Survival in San Francisco was significantly worse among people living in lower-income neighbourhoods. For Toronto though, no association was found between income and survival. Systemic health care issues, such as different health insurance coverage, were pointed out as the most plausible explanations for their findings. By contrast, still in the US, no evidence of racial (very much associated with SES in the US) inequalities were found within the Veterans Administration system in the US, a health care system with universal access [42]. Other studies found no association between socioeconomic condition and cancer outcome when comparing patients that had been offered treatment of the same type and same quality [43, 44]. In France, a small association was found between material deprivation and colorectal cancer survival [10]. However, the deprivation gap might have been overestimated since no deprivation-specific life tables were used. Other studies were inconclusive because they were based on overall survival or relative survival without deprivation-specific life tables [14, 37, 45]. Contrarily to these studies, we took in consideration the impact of plausible disparities in background mortality. The universal access nature of our healthcare system and the existence of a major public cancer reference centre which treats an important proportion of cancer patients of the north region could help explain the lack of association found between SES and survival. Nevertheless, further studies are needed to better understand between countries differences in the patients' pathway and healthcare organization that explain the existence or not of cancer survival inequalities.

Net survival was estimated in this study using the recently proposed estimator by Pohar-Perme [24]. This is an unbiased non-parametric estimator of the quantity of interest [46], when high quality information on cause of death is not available. Cancer data were provided by a population-based cancer registry (RORENO) that has been shown to have high completeness [47].

This study has some limitations that should be pointed out. We used area-based variables due to the absence of individual information. This can lead to some dilution of the effect. The education and the EDI levels attributed to each patient represent though the environment of his/her residence and not necessarily the individual condition. Furthermore, many other studies on the association between SES and survival from cancer have used ecological socioeconomic indices and still were able to find significant associations [11, 12, 14]. It has been shown that the size of the geographic unit is a key element for detecting inequalities [13]. The geographic unit we used to attribute the education level to each patient had a median population of 660 inhabitants, which correspond to a size comparable or lower than what has been used in those other similar studies. Another limitation of the study is the lack of information on stage of disease at diagnosis. Also information on comorbidities and treatment was not available.

Education level was measured as the proportion of individuals with at least nine years of education, i.e., the compulsory level of education in Portugal until recently. We have also used four years of education as cut-off, since this was the former compulsory level of education, and the results were similar (data not shown).

Patients analysed in this study were diagnosed in the period 2000-2002 which allowed for a long-term follow-up. These years correspond though to a period well before the economic crisis that began in 2008 and which affected Europe and particularly south European countries including Portugal. The National Health Service has been subject in recent years to budgetary constraints which may have led to inequalities in access to healthcare. Evaluations similar to the one presented in this study should be performed in the near future to access the impact of recent health policies in cancer survival inequalities. Other cancer sites should be analysed also to confirm, or not, the findings in this study.

The EDI is a recently developed indicator of socioeconomic deprivation. For Portugal the main variables used in the construction of this index were overcrowding, no indoor flushing, education, unemployment and not owning a house, reflecting though different domains of deprivation. Our study is one of the first studies to use this index. It would be interesting to compare SES inequalities in cancer survival across countries using this same index.

Conclusions

To the best of our knowledge, this is the first population-based study to address the question of socioeconomic inequalities in survival from colorectal cancer in Portugal. We found some inequalities in net survival by education level, but less by EDI, when using general life tables. However, the sensitivity analysis performed showed that these inequalities in cancer survival were most likely absent and were better explained by differences in background mortality. Our study confirms the importance of using the relevant life tables, or of performing sensitivity analysis, when evaluating socioeconomic inequalities in cancer survival.

Additional files

Additional file 1: Table S1. Age-standardized net survival estimates by education level and EDI. (DOCX 18 kb)

Additional file 2: Figure S1-S2. Figure S1 - Sensitivity analysis Education (Female Patients): Excess Hazard Ratios for the least educated group (compared with most educated group) at a) 5 years and b) 10 years since diagnosis. Figure S2 – Sensitivity analysis EDI (Female Patients): Excess Hazard Ratios for the most deprived group (compared with least deprived group) at a) 5 years and b) 10 years since diagnosis. (DOCX 92 kb)

Abbreviations

AIC, Akaike information criteria; CI, confidence interval; CRC, colorectal cancer; EDI, European deprivation index; EHR, excess hazard ratio; ICD, international classification of diseases; RORENO, North Region Cancer Registry of Portugal; SES, socioeconomic status; SNS, National Health Service ("Serviço Nacional de Saude")

Acknowledgments

The authors would like to thank Ana Isabel Ribeiro for her help in the matching process of patient's addresses and SES indicators.

Funding

This study was partially funded by a research grant from the Research Centre of Portuguese Oncology Institute – Porto (CI-IPOP 14-2012). BR was partly supported by Cancer Research UK (C1336-A11700).

Availability of data and materials

Cancer registry data can only be provided under special authorization from the Portuguese National Commission for Data Protection.

Authors' contributions

The author contributions were as follows: LA collected, performed the statistical analysis and interpreted the data, drafted and revised the manuscript. DM supervised the analysis and interpretation of data and reviewed the manuscript. MJB supplied the data, helped in the interpretation of the data and reviewed the manuscript. BR suggested the study, interpreted the data and reviewed the manuscript. All authors contributed to the discussion of the results. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interest.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The North Region Cancer Registry of Portugal (RORENO) is a population-based database of all patients who are diagnosed with cancer in the area covered by the registry. Routine demographic and clinical data are collected in accordance with the Portuguese privacy policy and with approval of the Portuguese Data Protection Authority. Individual patient consent is not required for the registry data. Patient anonymity is maintained by the coding of data during compilation and the access to information can be made available for research purposes since anonymized. Accordingly, all data released by RORENO for the current study was anonymized.

Author details

[1]Department of Epidemiology, Portuguese Oncology Institute (IPO Porto), Porto, Portugal. [2]RORENO - North Region Cancer Registry of Portugal, Porto, Portugal. [3]Faculty of Sciences, University of Porto, Porto, Portugal. [4]EPIUnit – Institute of Public Health – University of Porto (ISPUP), Porto, Portugal. [5]Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal. [6]UMIB, Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal. [7]Cancer Survival Group, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Received: 13 April 2016 Accepted: 27 July 2016
Published online: 05 August 2016

References

1. RORENO. Registo Oncológico Regional do Norte 2009. Porto: Instituto Português de Oncologia do Porto, Portugal; 2014.
2. Castro C, Antunes L, Lunet N, Bento MJ. Cancer incidence predictions in the North of Portugal: keeping population-based cancer registration up to date. Eur J Cancer Prev. 2015 (Epub ahead of print).
3. Ferlay J, Soerjomataram II, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman DD, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359–86.
4. De Angelis R, Sant M, Coleman MP, Francisci S, Baili P, Pierannunzio D, Trama A, Visser O, Brenner H, Ardanaz E, et al. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE–5-a population-based study. Lancet Oncol. 2014;15(1):23–34.
5. Manser CN, Bauerfeind P. Impact of socioeconomic status on incidence, mortality, and survival of colorectal cancer patients: a systematic review. Gastrointest Endosc. 2014;80(1):42–60.e49.
6. Aarts MJ, Lemmens VE, Louwman MW, Kunst AE, Coebergh JW. Socioeconomic status and changing inequalities in colorectal cancer? A review of the associations with risk, treatment and outcome. Eur J Cancer. 2010;46(15):2681–95.
7. Egeberg R, Halkjaer J, Rottmann N, Hansen L, Holten I. Social inequality and incidence of and survival from cancers of the colon and rectum in a population-based study in Denmark, 1994-2003. Eur J Cancer. 2008;44(14): 1978–88.
8. Hussain SK, Lenner P, Sundquist J, Hemminki K. Influence of education level on cancer survival in Sweden. Ann Oncol. 2008;19(1):156–62.
9. Cavalli-Björkman N, Lambe M, Eaker S, Sandin F, Glimelius B. Differences according to educational level in the management and survival of colorectal cancer in Sweden. Eur J Cancer. 2011;47(9):1398–406.
10. Dejardin O, Jones AP, Rachet B, Morris E, Bouvier V, Jooste V, Coombes E, Forman D, Bouvier AM, Launoy G. The influence of geographical access to health care and material deprivation on colorectal cancer survival: evidence from France and England. Health Place. 2014;30:36–44.
11. Rachet B, Ellis L, Maringe C, Chu T, Nur U, Quaresma M, Shah A, Walters S, Woods L, Forman D, et al. Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. Br J Cancer. 2010;103(4):446–53.
12. Ito Y, Nakaya T, Nakayama T, Miyashiro I, Ioka A, Tsukuma H, Rachet B. Socioeconomic inequalities in cancer survival: A population-based study of adult patients diagnosed in Osaka, Japan, during the period 1993-2004. Acta Oncol. 2014;53(10):1423–33.
13. Woods LM, Rachet B, Coleman MP. Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. Br J Cancer. 2005;92(7): 1279–82.
14. Dejardin O, Remontet L, Bouvier AM, Danzon A, Trétarre B, Delafosse P, Molinié F, Maarouf N, Velten M, Sauleau EA, et al. Socioeconomic and geographic determinants of survival of patients with digestive cancer in France. Br J Cancer. 2006;95(7):944–9.
15. Wrigley H, Roderick P, George S, Smith J, Mullee M, Goddard J. Inequalities in survival from colorectal cancer: a comparison of the impact of deprivation, treatment, and host factors on observed and cause specific survival. J Epidemiol Community Health. 2003;57(4):301–9.
16. Guillaume E, Pornet C, Dejardin O, Launay L, Lillini R, Vercelli M, Marí-Dell'Olmo M, Fernández Fontelo A, Borrell C, Ribeiro AI, et al. Development of a cross-cultural deprivation index in five European countries. J Epidemiol Community Health. 2016;70(5):493–9.
17. Pornet C, Delpierre C, Dejardin O, Grosclaude P, Launay L, Guittet L, Lang T, Launoy G. Construction of an adaptable European transnational ecological deprivation index: the French version. J Epidemiol Community Health. 2012; 66(11):982–9.
18. Frederiksen BL, Osler M, Harling H, Ladelund S, Jørgensen T. Do patient characteristics, disease, or treatment explain social inequality in survival from colorectal cancer? Soc Sci Med. 2009;69(7):1107–15.
19. Pinto CG, Paquete A, Pissarra I. Colorectal cancer in Portugal. Eur J Health Econ. 2010;10:65–73.
20. Ferlay J, Burkhard C, Whelan S, Parkin DM. Check and conversion programs for cancer registries, vol. IARC Technical Report No. 42. Lyon: IARC; 2005.
21. World Health Organization. International Statistical Classification of Diseases and Related Health Problems, 10th revision vol. 2. Geneva: WHO; 2010.
22. GPSVisualizer [http://www.gpsvisualizer.com/geocoder/]. Accessed 29 July 2016.
23. INE. Antecedentes, metodologia e conceitos: Censos 2001. Lisboa: Instituto Nacional de Estatística; 2003.

24. Perme MP, Stare J, Estève J. On estimation in relative survival. Biometrics. 2012;68(1):113–20.
25. Spika D, Rachet B, Bannon F, Woods LM, Maringe C, Bonaventure A, Coleman MP, Allemani C: Life tables for the CONCORD-2 study. Available from: http://csg.lshtm.ac.uk/life-tables, downloaded on 13 December 2014.
26. Rachet B, Maringe C, Woods LM, Ellis L, Spika D, Allemani C. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. BMC Public Health. 2015;15:1240.
27. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. Stat Med. 2007; 26(30):5486–98.
28. Clerc-Urmès I, Grzebyk M, Hédelin G. Net survival estimation with stns. Stata J. 2014;14(1):87–102.
29. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata J. 2009;9(2):265–90.
30. Auvinen A, Karjalainen S. Possible explanations for social class differences in cancer patient survival. IARC Sci Publ. 1997;138:377–97.
31. Statistics Portugal [http://www.ine.pt]. Accessed 29 July 2016.
32. Ribeiro S, Furtado C, Pereira J. Association between cardiovascular disease and socioeconomic level in Portugal. Rev Port Cardiol. 2013;32(11):847–54.
33. Nogueira H, Remoaldo P. Olhares Geográficos sobre a Saúde. 1st ed. Lisboa: Edições Colibri; 2010.
34. Direção-Geral da Saúde. Programa Nacional para as Doenças Oncológicas - Relatório 2014, Avaliação e Monitorização dos Rastreios Oncológicos Organizados de Base Populacional de Portugal Continental. Lisboa: Direcão-Geral da Saúde; 2015.
35. Forno SEA, Poças FC, Matos ME. O cancro colorretal e o rastreio: conhecimentos e atitudes dos portuenses. Portuguese Journal of Gastroenterology. 2012;19:118–25.
36. Gorey KM, Luginaah IN, Bartfay E, Fung KY, Holowaty EJ, Wright FC, Hamm C, Kanjeekal SM. Effects of socioeconomic status on colon cancer treatment accessibility and survival in Toronto, Ontario, and San Francisco, California, 1996-2006. Am J Public Health. 2011;101(1):112–9.
37. Harris AR, Bowley DM, Stannard A, Kurrimboccus S, Geh JI, Karandikar S. Socioeconomic deprivation adversely affects survival of patients with rectal cancer. Br J Surg. 2009;96(7):763–8.
38. Kim J, Artinyan A, Mailey B, Christopher S, Lee W, McKenzie S, Chen SL, Bhatia S, Pigazzi A, Garcia-Aguilar J. An interaction of race and ethnicity with socioeconomic status in rectal cancer outcomes. Ann Surg. 2011;253(4):647–54.
39. Mackillop WJ, Zhang-Salomons J, Groome PA, Paszat L, Holowaty E. Socioeconomic status and cancer survival in Ontario. J Clin Oncol. 1997; 15(4):1680–9.
40. Gorey KM, Holowaty EJ, Fehringer G, Laukkanen E, Moskowitz A, Webster DJ, Richter NL. An international comparison of cancer survival: Toronto, Ontario, and Detroit, Michigan, metropolitan areas. Am J Public Health. 1997;87(7): 1156–63.
41. Mitry E, Rachet B, Quinn MJ, Cooper N, Coleman MP. Survival from cancer of the rectum in England and Wales up to 2001. Br J Cancer. 2008;99 Suppl 1:S30–2.
42. Rabeneck L, Souchek J, El-Serag HB. Survival of colorectal cancer patients hospitalized in the Veterans Affairs Health Care System. Am J Gastroenterol. 2003;98(5):1186–92.
43. Lyratzopoulos G, Sheridan GF, Michie HR, McElduff P, Hobbiss JH. Absence of socioeconomic variation in survival from colorectal cancer in patients receiving surgical treatment in one health district: cohort study. Colorectal Dis. 2004;6(6):512–7.
44. Nur U, Rachet B, Parmar MK, Sydes MR, Cooper N, Lepage C, Northover JM, James R, Coleman MP, Collaborators A. No socioeconomic inequalities in colorectal cancer survival within a randomised clinical trial. Br J Cancer. 2008;99(11):1923–8.
45. Pollock AM, Vickers N. Breast, lung and colorectal cancer incidence and survival in South Thames Region, 1987-1992: the effect of social deprivation. J Public Health Med. 1997;19(3):288–94.
46. Roche L, Danieli C, Belot A, Grosclaude P, Bouvier AM, Velten M, Iwaz J, Remontet L, Bossard N. Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. Int J Cancer. 2013;132(10):2359–69.
47. Castro C, Bento MJ, Lunet N, Campos P. Assessing the completeness of cancer registration using suboptimal death certificate information. Eur J Cancer Prev. 2012;21(5):478–9.

Table S1 – Age-standardized net survival estimates by education level and EDI

	Male						Female					
	1-year		5-years		10-years		1-year		5-years		10-years	
	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Education level												
Higher education	80	76 - 84	59	54 - 64	55	48 - 63	81	77 - 85	58	53 - 63	55	49 - 61
q4	79	75 - 82	58	52 - 63	51	44 - 58	82	79 - 86	53	48 - 58	49	43 - 55
q3	83	79 - 86	56	50 - 61	48	41 - 55	82	78 - 86	57	51 - 63	54	47 - 61
q2	82	77 - 86	53	47 - 58	42	36 - 48	85	81 - 89	65	59 - 71	55	48 - 63
Lower education	82	78 - 86	54	48 - 60	42	36 - 49	74	68 - 79	51	44 - 58	46	38 - 53
EDI												
Least deprived	80	76 - 85	58	52 - 64	52	44 - 60	82	78 - 87	59	52 - 65	55	48 - 63
q4	80	75 - 84	56	51 - 62	57	48 - 65	78	73 - 83	59	53 - 65	52	45 - 59
q3	81	77 - 85	58	52 - 63	45	39 - 51	82	77 - 86	55	49 - 61	48	41 - 55
q2	79	75 - 83	54	49 - 59	44	37 - 51	84	80 - 88	56	50 - 61	52	45 - 58
Most deprived	83	80 - 86	55	50 - 59	48	41 - 54	80	77 - 84	56	51 - 61	53	47 - 59

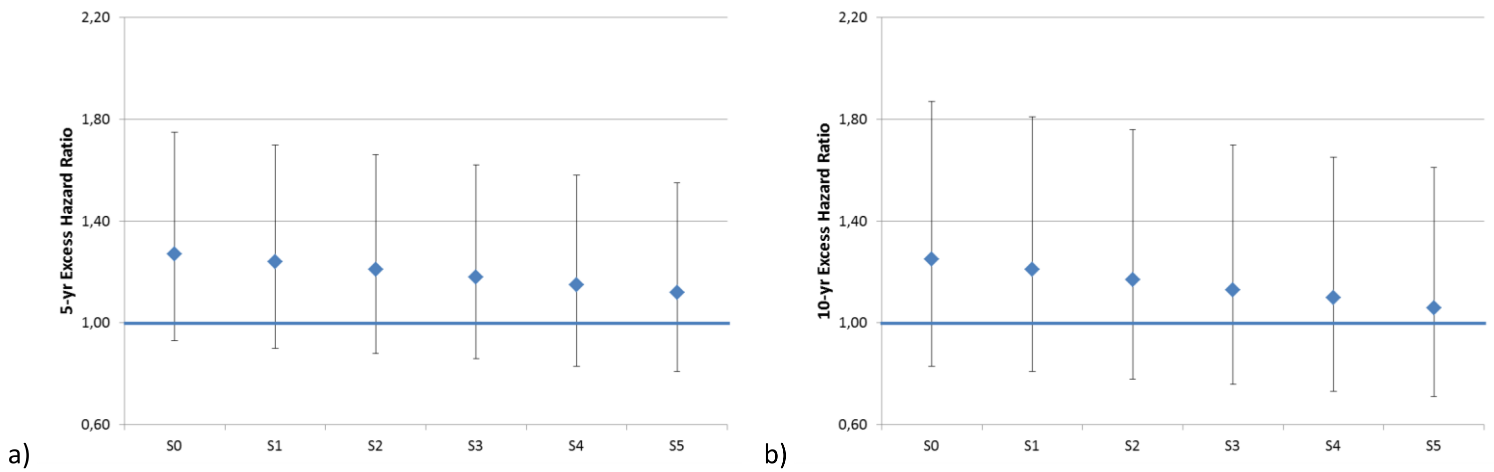

Figure S1 – Sensitivity analysis Education (Female Patients): Excess Hazard Ratios for the least educated group (compared with most educated group) at a) 5 years and b) 10 years since diagnosis.

S0 – general life tables
S1 – education-specific life tables with 20% of the English gap between education groups
S2 - education-specific life tables with 40% of the English gap between education groups
S3 - education-specific life tables with 60% of the English gap between education groups
S4 - education-specific life tables with 80% of the English gap between education groups
S5 - education-specific life tables with the English gap between education groups

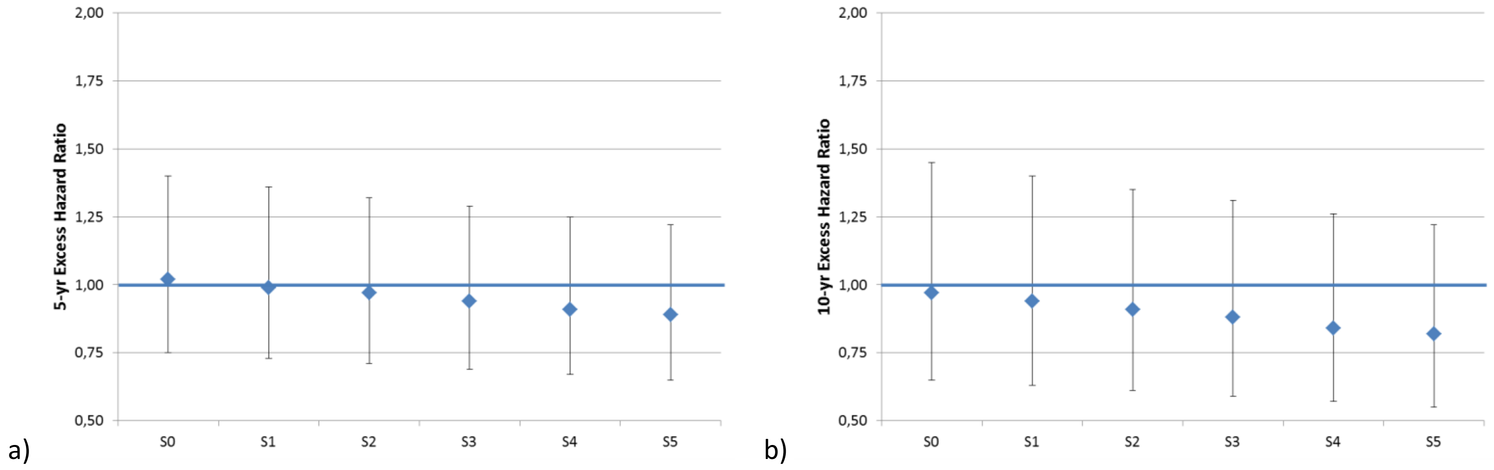

Figure S2 – Sensitivity analysis EDI (Female Patients): Excess Hazard Ratios for the most deprived group (compared with least deprived group) at a) 5 years and b) 10 years since diagnosis.

S0 – general life tables
S1 – EDI-specific life tables with 20% of the English gap between deprivation groups
S2 - EDI-specific life tables with 40% of the English gap between deprivation groups
S3 - EDI-specific life tables with 60% of the English gap between deprivation groups
S4 - EDI-specific life tables with 80% of the English gap between deprivation groups
S5 - EDI-specific life tables with the English gap between deprivation groups

3.3 Study III: Building deprivation-specific life tables

Deprivation-specific life tables using multivariable flexible modelling - trends from 2000-2002 to 2010-2012, Portugal

Luís Antunes, Denisa Mendonça, Ana Isabel Ribeiro, Camille Maringe, Bernard Rachet

(Submitted)

The correct evaluation of net survival by deprivation group demands using deprivation-specific life tables to correctly adjust for the background mortality. The lack of these tables for Portugal led to the need of building them. Mortality and population information, necessary to accomplish this objective, were obtained from the Statistics Portugal officce (INE). The smallest area for which this information was available, stratified by age group and sex was parish. All parishes were classified in deprivation quintiles according to the European Deprivation Index developed for Portugal. Number of deaths and population of all parishes in the same quintile were summed up. The mortality rates were then modelled using a multivariable flexible Poisson model. The age dependence was modelled using cubic regression splines. Besides age, deprivation group, calendar period and interactions age*deprivation, age*period and deprivation*period were considered in the model. Men and women were modelled separately. Using the predicted mortality rates from the model, the life tables were built and life expectancy at birth, stratified by age, sex, deprivation group and calendar period. Mortality rate ratios and corresponding confidence intervals were obtained from the fitted models. The formulas for calculating the variance of the model-based predicted mortality rate ratios were derived.

Persistent differences in mortality and life expectancy were observed according to ecological socioeconomic deprivation. The differences were larger among men and decreased with age for both sexes. Mortality decreased significantly between the two periods leading to an improvement in life expectancy between two and three years. Deprivation gaps in mortality/life expectancy remained nearly constant over the ten-year period analysed.

Next, the resulting manuscript of this study is presented.

Deprivation-specific life tables using multivariable flexible modelling – trends from 2000-2002 to 2010-2012, Portugal

Luís Antunes[1,2,3] – luis.antunes@ipoporto.min-saude.pt

Denisa Mendonça[3,4] – dvmendon@icbas.up.pt

Ana Isabel Ribeiro[3,5] - ana.isabel.ribeiro@ispup.up.pt

Camille Maringe[6] – camille.maringe@lshtm.ac.uk

Bernard Rachet[6] – bernard.rachet@lshtm.ac.uk

[1] Grupo de Epidemiologia do Cancro, Centro de Investigação do IPO Porto (CI-IPOP), Instituto Português de Oncologia do Porto (IPO Porto), Porto, Portugal

[2] Faculdade de Ciências, Universidade do Porto, Portugal

[3] EPIUnit - Instituto de Saúde Pública, Universidade do Porto, Rua das Taipas, nº 135, 4050-600 Porto, Portugal

[4] Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Portugal

[5] Departamento de Ciências da Saúde Pública e Forenses e Educação Médica, Faculdade de Medicina, Universidade do Porto, Porto, Portugal.

[6] Cancer Survival Group, London School of Hygiene and Tropical Medicine, United Kingdom

Corresponding author:

Luis Antunes

Rua Dr. António Bernardino de Almeida

4200-072 Porto, Portugal

Email: luis.antunes@ipoporto.min-saude.pt

Abstract

Background: Completing mortality data by information on possible socioeconomic inequalities in mortality is crucial for policy planning. The aim of this study was to build deprivation-specific life tables using the Portuguese version of the European Deprivation Index (EDI) as a measure of area-level socioeconomic deprivation, and to evaluate mortality trends between the periods 2000-2002 and 2010-2012.

Methods: Statistics Portugal provided the counts of deaths and population by sex, age group, calendar year and area of residence (parish). A socioeconomic deprivation level was assigned to each parish according to the quintile of their national EDI distribution. Death counts were modelled within the generalised linear model framework as a function of age, deprivation level and calendar period. Mortality Rate Ratios (MRR) were estimated to evaluate variations in mortality between deprivation groups and periods.

Results: Life expectancy at birth increased from 74.0 and 80.9 years in 2000-2002, for men and women, respectively, to 77.6 and 83.8 years in 2010-2012. Yet, life expectancy at birth differed by deprivation, with, compared to least deprived population, a deficit of about 2 (men) and 1 (women) years among most deprived in the whole study period. The higher mortality experienced by most deprived groups at birth (in 2010-2012, mortality rate ratios of 1.74 and 1.29 in men and women, respectively) progressively disappeared with increasing age.

Conclusions: Persistent differences in mortality and life expectancy were observed according to ecological socioeconomic deprivation. These differences were larger among men and mostly marked at birth for both sexes.

Keywords: life-tables, deprivation, multivariable modelling, socioeconomic factors, health inequalities.

Introduction

Life tables provide information on mortality rates and probabilities of death for specific populations defined by geographical regions and/or periods of time. They are important demographic tools as they are the basis for the estimation of life expectancy at birth, an important indicator of population health and development. Many factors are known to influence overall mortality, such as age, sex, geographical region, socioeconomic deprivation or ethnicity [1–4]. While the effect of, for example, age is largely unavoidable, the gap in mortality due to socioeconomic characteristics could be reduced with policies oriented to improve population living conditions and to change the social and economic structures [5].

Many studies showed the existence of socioeconomic inequalities in health outcomes including mortality [6–9]. These inequalities can result from different lifestyle behaviours, namely, smoking, alcohol, physical activity and dietary habits, different health literacy or access to health care, among other factors. However, no life tables have been constructed by socioeconomic level in Portugal yet. An ecological measure of socioeconomic deprivation, the European Deprivation Index (EDI) [10,11] has recently become available in Portugal. The first aim of this study was thus to build deprivation-specific life tables using the Portuguese version of the EDI. The second aim was to evaluate mortality ratios between deprivation groups and trends in inequalities between 2000-2002 and 2010-2012 in Portugal.

Methods

Socioeconomic deprivation

The Portuguese version of the European Deprivation Index was used as deprivation indicator. This index was built using a methodology first proposed by Pornet and colleagues in 2012 [12] and then applied to several European countries including Portugal [10,11]. The index is based on census variables available for each country that are most associated with variables identified from the European Union Statistics on Income and Living Conditions (EU-SILC) survey [13]. The index for Portugal based on 2001 census includes percentage of: non-owned households, households without indoor flushing, residents with low education level (≤6th grade), household with 5 rooms or less, unemployed looking for a job, female residents aged 65 years or more, households without bath/shower and percentage of residents employed in manual occupations [11]. A score was obtained for each parish based on the census responses of its inhabitants. This score was then categorized in five quintiles from the least deprived (q1) to the most deprived (q5) such that each quintile corresponded to 20% of the Portuguese population. Each deceased was assigned with the deprivation quintile corresponding to his/her parish of residence at the time of death.

Mortality and population data

Mortality rates in life tables require counts of deaths (numerator of the rates) and population (denominator) stratified by demographic variables (age, sex, others). This information is usually made available by the national statistics offices. The number of deaths by sex, age group (0, 1-4, 5-9, …, 85+), year of death and area of residence (parish) was obtained by special request to the Statistics Portugal (*Instituto Nacional de Estatística*). As common practice, to increase estimates' stability, three years of data were considered centred on each census year for which the life tables were estimated (2000-2002 and 2010-2012). Population data was retrieved from the Statistics Portugal website (www.ine.pt). Number of residents by sex, age group (0, 1-4, 5-9, …, 85+) and parish was only available for census years (2001, 2011) so that the population was considered constant over the three years of each studied period. There were 4,241 parishes in Portugal in 2001, with a median population of 969 inhabitants (min-max: 39-81,845), while in 2011 this number increased to 4,260 (median population: 892, min-max: 31-66,250).

In this study, both the numbers of deaths and people (residing in the parishes) were summed up across the parishes for each period by sex, age group and level of deprivation.

Statistical analysis

When at subnational level, the counts of deaths and population produced by the national statistics offices are often available only by age groups (e.g. abridged) rather than by single years of age (e.g. complete) [14]. Several methods for building complete life tables from abridged data have been in use, namely, Elandt–Johnson, Kostaki, Brass logit, and Akima spline methods [15]. More recently, Rachet and colleagues [14] suggested a modelling approach to estimate smoothed mortality rates using flexible Poisson multivariable models. Death counts are modelled in the generalised linear model framework, considering a Poisson error and using splines to capture the effect of age. This method can use complete or abridged raw data allowing the estimation of complete life tables. This type of models was considered recommendable because it derives robust and unbiased estimates without making strong assumptions about age-specific mortality profiles. Also, a simulation study has shown that this method had better goodness of fit performance than other implemented methods [14]. The age-group specific death counts were here modelled within this generalised linear model framework, considering a Poisson error with a log link function. The offset was considered the person-years at risk. Male and female death counts were modelled separately. Covariates considered in the model were age (using the mid-age of each age group), quintile of deprivation (*dep*), period (2000-2002 vs 2010-2012) and interactions between deprivation and age, deprivation and period and period and age. The model can be written as:

$$\log\left(d_{age,i,j}\right) = \beta_0 + f(age) + \sum_{i=2}^{5} \beta_{d_i} \cdot dep_i + g_1(age * dep_i)$$

$$+ \beta_{per_j} \cdot period_j + g_2(age * period_j) + \sum_{k=2}^{5} \beta_{dep\_per_k} \cdot dep_k * period_j + \log\left(pyrs_{age,i,j}\right),$$

where $d_{age,i,j}$ denotes the number of deaths and $pyrs_{age,i,j}$ the number of person-years at risk for each age, deprivation group *i* and period *j*. The functions *f*, g_1 and g_2 represent restricted cubic splines. The knots positions were fixed *a priori* at ages 0, 1, 2 and 88 (for men) or 89 (for women). Although Rachet and colleagues considered further five knot positions selected from a set of 100 randomly simulated locations, here, we opted to consider knots at ages 10 to 50 at 10 years intervals since the other approach produced unrealistic predicted values. From these predefined knots position, the final number of knots was selected based on the Akaike Information Criterion (AIC).

Mortality rates were predicted from the fitted models by individual year of age (0-99), for each sex, period and quintile of deprivation. Life expectancies at birth were calculated from the fitted life tables. Mortality rate ratios (MRR) in terms of age were calculated from the predicted mortalities. MRR by EDI were calculated using the least deprived group as reference and the MRR by period using the period 2000-2002 as reference. The 95% confidence intervals (CI) for MRR were built assuming a normal distribution of log MRR and using the delta method. The derivations of the expressions for the CIs are presented as Supplementary Material (S1).

All calculations were performed using STATA v13.1 and R v3.4.0. The STATA command *mvrs* was used for fitting the flexible Poisson model [16].

Results

In the period 2000-2002 a total of 316,714 deaths were observed from which 219 (0.07%) were excluded due to unknown parish of residence at the time of death. In the period 2010-2012, the total number of deaths was 316,410 and only 75 deaths (0.02%) were excluded due to unknown parish or unknown age.

Both deprivation and period were found statistically significantly associated with mortality and all final fitted models included interactions between age and deprivation, period and age and period and deprivation, which were also found to be statistically significant. Higher mortality rates increasing with higher deprivation levels and decreasing with time periods were observed. The predicted mortality rates by age, sex, period and deprivation quintiles are presented in Supplementary Tables S1 to S4 and in Figure 1 and Figure 2. For all combinations period-sex-EDI analysed, the mortality rate first decreased with age reaching a minimum around 8-10 years-old and then steadily increased with age. For all deprivation quintiles q2 to q5, the mortality rates were in general significantly higher than the mortality rates of the least deprived group (q1) (Figure 3). In men, at birth, the MRR between the most and the least deprived group was 1.62 (95%CI: 1.54-1.70) and 1.74 (95%CI: 1.65-1.83) in periods 2000-2002 and 2010-2012, respectively. The MRRs decreased with age and from age 81 onwards the ratio was no longer significantly different from 1 in the first period and from age 93 onwards for the most recent period (Figure 3, top). In women, the MRRs were lower than the ones observed for men: at birth, the MRR between the two extreme deprivation groups was 1.26 (1.18-1.35) and 1.29 (1.20-1.38) in 2000-2002 and 2010-2012, respectively. The MRRs decreased with age but remained always significantly above one (Figure 3, bottom). A reduction in the mortality rates was observed from the period 2000-2002 to the most recent period, 2010-2012. This reduction was significant for all ages (Figure 4) and for both men and women.

However, it was observed that the relative decrease in mortality was between 40% and 60% in less than 40 year olds but only 20% in older ages (over 60 year old) in men. This age gradient in the relative decrease between the two periods is less marked in women. For men, the reduction of mortality over time was less favourable for the most deprived group.

Life expectancy at birth increased from 74.0 years in 2000-2002 to 77.6 years in 2010-2012 in men and from 80.9 to 83.8 years in women. The gap in life expectancy at birth between the least and the most deprived group for men in the first period was 1.8 years. This gap slightly increased, over the ten-year period, to 2.1 years. For women, a smaller gap in life expectancy compared to men was observed. In 2000-2002 it was 1.0 year and it remained almost nearly constant over time (0.9 years in 2010-2012). The gap in life expectancy at 65 years was lower than at birth in both sexes. For men: 0.3 and 0.7 years for 2000-2002 and 2010-2012, respectively. In women, it was 0.5 years for both periods.

Discussion

Persistent differences in mortality and life expectancy were observed according to ecological socioeconomic deprivation. These differences were larger among men and decreased with age for both sexes. Although mortality decreased significantly between the two periods, the deprivation gaps in mortality/life expectancy remained nearly constant from the period 2000-2002 to 2010-2012. The smaller socioeconomic inequalities in mortality found in women have also been observed in other countries [17–19]. Several factors can contribute for this different pattern, including health behaviours and occupation. According to the last national health surveys made in Portugal, the prevalence of smoking was higher in men with low socioeconomic status while in women the prevalence was higher among individuals of high socioeconomic status [20]. Ribeiro and colleagues [21] found no influence of deprivation on longevity after 75 for men and a weak association for women in Portugal. We observed here a decrease in the mortality rate ratios between deprivation groups with age. Similarly to Ribeiro and colleagues findings, the difference between mortality rates ceased to be significant after ages around 80 years in men, while in women a slight but still significant difference remained at all ages.

Richardson and colleagues analysed the evolution in regional gap in life expectancy at birth from 1991 to 2008 within European Union countries [22]. No reduction in life expectancy gaps over the two decades analysed was observed, similarly to what has been observed in this study.

EUROSTAT publishes estimates of life expectancy by age, sex and educational attainment level for several countries of the European Union including Portugal [23]. The difference in the estimates of life expectancy at birth between the two extreme education groups, "Less than primary", "primary and lower secondary education" and "Tertiary education", presents a high variability between countries. While this difference, for males in 2011, was 19.3 years in Czech Republic, it was as low as 3.6 years in Turkey. In Portugal, this gap was 4.5 years, the third lowest within the 17 EU countries with published information. For women, the gaps were generally lower, ranging from 1.7 (Italy and Malta) to 8.7 years (Bulgaria). In Portugal it was 2.0 years. The smaller gap for women is in accordance with our findings. While we found smaller gaps between extreme deprivation groups, this can be justified by our use of an ecological indicator versus the individual information on education used by EUROSTAT. The use of an ecological index can introduce some dilution effect due to the socioeconomic heterogeneity within the geographical regions considered. Nevertheless, we used the smallest geographical area for which mortality data stratified by age groups and sex are available.

In this study we used a modelling approach to predict the mortality rate profiles by age. The flexible Poisson models have already been shown useful and valid to build life tables. They were used to build region-specific life tables within the CONCORD study [24] and by region, deprivation and ethnicity in England [4,25] . The multivariable model allowed incorporating the deprivation and period effects, as well as interactions between them and with age, in a single model. Also it allowed obtaining complete life tables from abridged raw data.

Inference based on model predictions must have in consideration the correlation between the estimated parameters of the model. Ignoring this dependence and using the classical variance formulas as if the predicted values were observed ones would result in an underestimation of the confidence intervals range. We thus derived and presented the variance estimators for the model-based mortality rate ratios taking into account this dependence.

This study is very relevant for the surveillance and monitoring of health inequalities, but it is important to highlight that these specific life tables are crucial tools to obtain reliable estimates of cancer survival within the relative survival data setting. In this setting, information on the cause of death is not available or not reliable. The disease-related survival (net survival) is then obtained indirectly by comparing the all-cause mortality of the cohort of patients with the mortality that would be experienced by individuals with the same demographic characteristics but free of the disease [26]. The information on this expected (also called background) mortality is obtained from population life tables, assuming that the mortality due to the disease in question is negligible relatively to the overall mortality [27]. To obtain valid net survival estimates, the population mortality should correctly reflect the expected mortality for each patient. The use of general life tables in the estimation of net survival for subgroups of the population with different overall mortality can lead to biased estimates of net survival. Estimation of net survival by deprivation is a situation where the use of general life tables can lead to overestimation of net survival in affluent groups, if these groups have a lower overall mortality than the general population, and the underestimation of net survival in the deprived groups, if these groups have a higher mortality than the general population [28,29].

These questions arise also when stratifying by other factors that can influence overall mortality such as ethnicity [30].

Conclusion

In conclusion, this study has shown the existence of persistent socioeconomic inequalities in overall mortality in Portugal. Deprivation-specific life tables were built for Portugal. These life tables can therefore be used for monitoring inequalities and in future studies that require background mortality information in the estimation of deprivation-specific net survival from any specific disease.

List of abbreviations

CI – Confidence interval

EDI – European Deprivation Index

MRR – Mortality Rate Ratio

Declarations

Ethics approval and consent to participate

Only aggregated data publicly available from the Statistics Portugal office was used. No ethics approval was needed for this study.

Consent for publication

Not applicable.

Availability of data and material

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

None declared.

Funding

Ana Isabel Ribeiro was supported during this study by the European Commission (Horizon 2020 grant number 633666) and by FEDER through the Operational Programme Competitiveness and Internationalization and national funding from the Foundation for Science and Technology – FCT (Portuguese Ministry of Science, Technology and Higher Education) under the Unidade de Investigação em Epidemiologia - Instituto de Saúde Pública da Universidade do Porto (EPIUnit) (POCI-01-0145-FEDER-006862; Ref. UID/DTP/04750/2013). Camille Maringe was supported by Cancer Research UK grant number C7923/A18525.

Authors' contributions

The author contributions were as follows: LA designed the study, performed the statistical analysis and wrote the manuscript. DM supervised the analysis and interpretation of data and reviewed the manuscript. AIR and CM contributed to the statistical analysis and reviewed the manuscript. BR interpreted the data and reviewed the manuscript. All authors contributed to the discussion of the results. All authors read and approved the final version of the manuscript.

Acknowledgements

Not applicable.

Conflicts of interest

None declared.

References

1 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;59:115–20.

2 Auger N, Alix C, Zang G, Daniel M. Sex, age, deprivation and patterns in life expectancy in Quebec, Canada: A population-based study. *BMC Public Health* 2010;10. doi:10.1186/1471-2458-10-161.

3 Tobias MI, Cheung J. Monitoring health inequalities: life expectancy and small area deprivation in New Zealand. *Popul Health Metr* 2003;1:2.

4 Morris M, Woods LM, Rachet B. A novel ecological methodology for constructing ethnic-majority life tables in the absence of individual ethnicity information. *J Epidemiol Community Heal* 2015;69:361–67.

5 Mackenbach JP, Bakker MJ. Tackling socioeconomic inequalities in health: analysis of European experiences. *Lancet* 2003;362:1409–14.

6 Pappas G, Queen S, Hadden W, Fisher G. The Increasing Disparity in Mortality between Socioeconomic Groups in the United States, 1960 and 1986. *N Engl J Med* 1993;329:103–9.

7 Mackenbach JP, Kunst AE. Measuring the magnitude of socio-economic inequalities in health: An overview of available measures illustrated with two examples from Europe. *Soc Sci Med* 1997;44:757–71.

8 van Lenthe FJ, Borrell LN, Costa G, et al. Neighbourhood unemployment and all cause

mortality: a comparison of six countries. *J Epidemiol Community Heal* 2005;59:231–37.

9 Kohler I V, Martikainen P, Smith KP, Elo IT. Educational differences in all-cause mortality by marital status - Evidence from Bulgaria, Finland and the United States. *Demogr Res* 2008;19:2011–42.

10 Guillaume E, Pornet C, Dejardin O, et al. Development of a cross-cultural deprivation index in five European countries. *J Epidemiol Community Health* 2015;jech-2015-205729.

11 Ribeiro AI, Mayer A, Miranda A, De Pina M de F. The Portuguese Version of the European Deprivation Index: An Instrument to Study Health Inequalities. *Acta Med Port* 2017;30:17.

12 Pornet C, Delpierre C, Dejardin O, et al. Construction of an adaptable European transnational ecological deprivation index: the French version. *J Epidemiol Community Health* 2012;66:982–89.

13 Eurostat. Access to Microdata. EUROPEAN UNION STATISTICS ON INCOME AND LIVING CONDITIONS (EU-SILC). 2015.URL http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions Accessed 19 May 2018.

14 Rachet B, Maringe C, Woods LM, Ellis L, Spika D, Allemani C. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health* 2015;15:1240.

15 Baili P, Micheli A, Montanari A, Capocaccia R. Comparison of four methods for estimating complete life tables from abridged life tables using mortality data supplied to eurocare-3. *Math Popul Stud* 2005;12:183–98.

16 Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: A principled approach. *Stata J* 2007;7:45–70.

17 Stronks K, van de Mheen H, van den Bos J, Mackenbach JP. Smaller socioeconomic inequalities in health among women: the role of employment status. *Int J Epidemiol*

1995;24:559–68.

18 Mackenbach JP, Kunst AE, Groenhof F, et al. Socioeconomic inequalities in mortality among women and among men: an international study. *Am J Public Health* 1999;89:1800–1806.

19 Mustard CA, Etches J. Gender differences in socioeconomic inequality in mortality. *J Epidemiol Community Health* 2003;57:974–80.

20 Alves J, Kunst AE, Perelman J. Evolution of socioeconomic inequalities in smoking: results from the Portuguese national health interview surveys. *BMC Public Health* 2015;15:311.

21 Ribeiro AI, Krainski ET, Sá Carvalho M, Launoy G, Pornet C, Pina M de F de. Does community deprivation determine longevity after the age of 75? A cross-national analysis. *Int J Public Health* 2018;0123456789. doi:10.1007/s00038-018-1081-y.

22 Richardson EA, Pearce J, Mitchell R, Shortt NK, Tunstall H. Have regional inequalities in life expectancy widened within the European Union between 1991 and 2008? *Eur J Public Health* 2014;24:357–63.

23 EUROSTAT. Life expectancy by age, sex and educational attainment level. 2017.URL http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do Accessed 12 May 2018.

24 Spika D, Bannon F, Bonaventure A, et al. Life tables for global surveillance of cancer survival (the CONCORD programme): data sources and methods. *BMC Cancer* 2017;17:159.

25 Cancer Research UK Cancer Survival Group. UK life tables. URL http://csg.lshtm.ac.uk/tools-analysis/uk-life-tables/ Accessed 19 May 2018.

26 Perme MP, Stare J, Estève J. On Estimation in Relative Survival. *Biometrics* 2012;68:113–20.

27 Esteve J, Benhamou E, Raymond L. Techniques for survival analysis Survival analysis in descriptive epidemiology. In: *Statistical Methods in Cancer Research Volume IV - Descriptive Epidemiology*. Lyon: IARC, 1994.

28 Auvinen a, Karjalainen S. Possible explanations for social class differences in cancer patient

survival. *IARC Sci Publ* 1997;377–97.

29 Dickman PW, Auvinen A, Voutilainen ET, Hakulinen T. Measuring social class differences in cancer patient survival: is it necessary to control for social class differences in general population mortality? A Finnish population-based study. *J Epidemiol Community Health* 1998;52:727–34.

30 Blakely T, Soeberg M, Carter K, Costilla R, Atkinson J, Sarfati D. Bias in relative survival methods when using incorrect life-tables: Lung and bladder cancer by smoking status and ethnicity in New Zealand. *Int J Cancer* 2012;131:974–82.

Figures

Figure 1 – Predicted mortality rates (log scale) according to quintiles of socioeconomic deprivation 2000-2002 for Men (top) and Women (bottom).

Figure 2 – Predicted mortality rates (log scale) according to quintiles of socioeconomic deprivation 2010-2012 for Men (top) and Women (bottom).

Figure 3 – Mortality Rate Ratio as function of age between deprivation quintiles q2, q3, q4 and q5 and least deprived quintile (q1) for Men (top) and Women (bottom).

Figure 4 – Mortality Rate Ratio as function of age between period 2010-2012 and 2000-2002 (reference) according to deprivation quintile for men (left) and women (right).

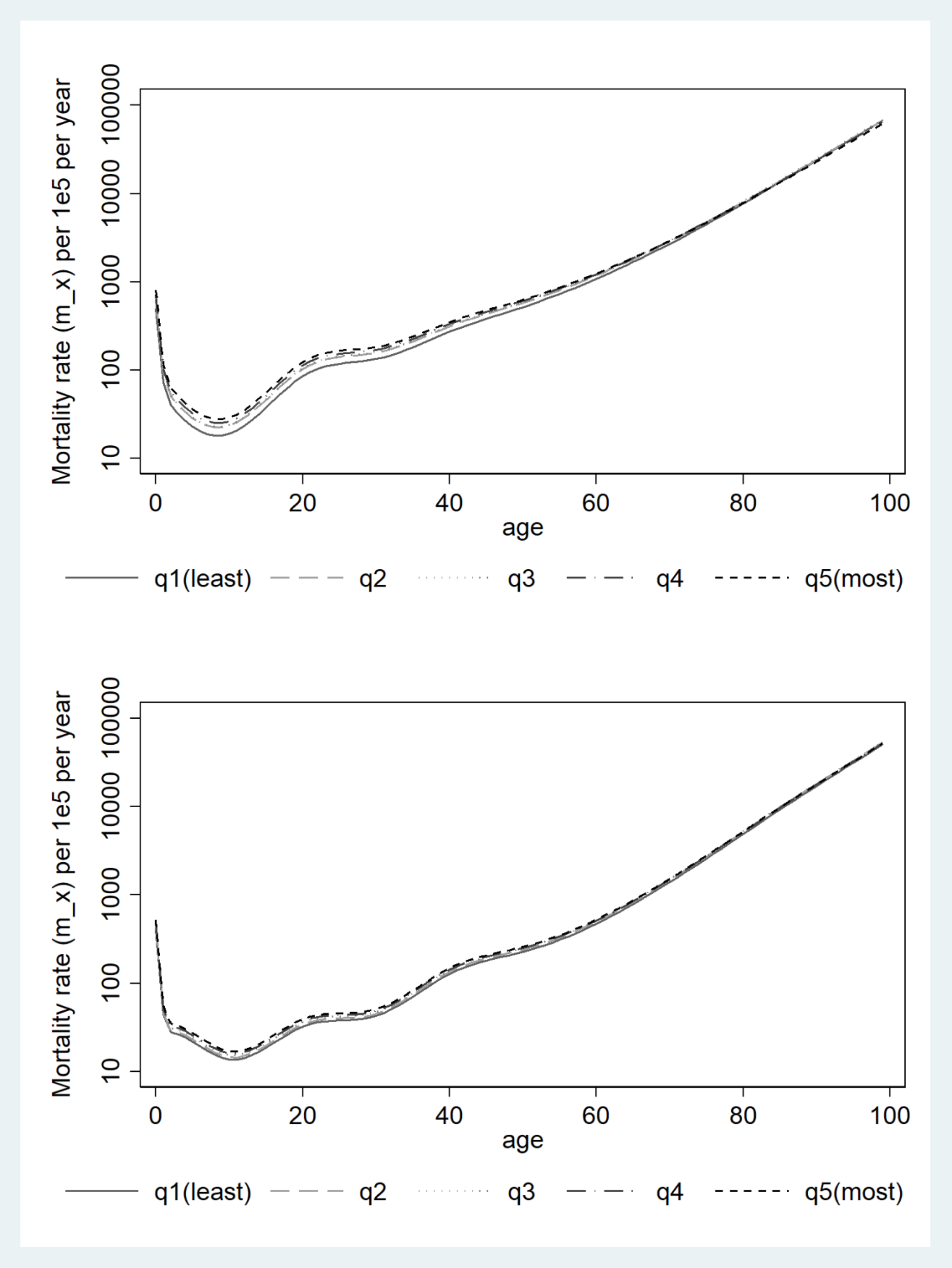

Figure 1 – Predicted mortality rates (log scale) according to quintiles of socioeconomic deprivation 2000-2002 for Men (top) and Women (bottom).

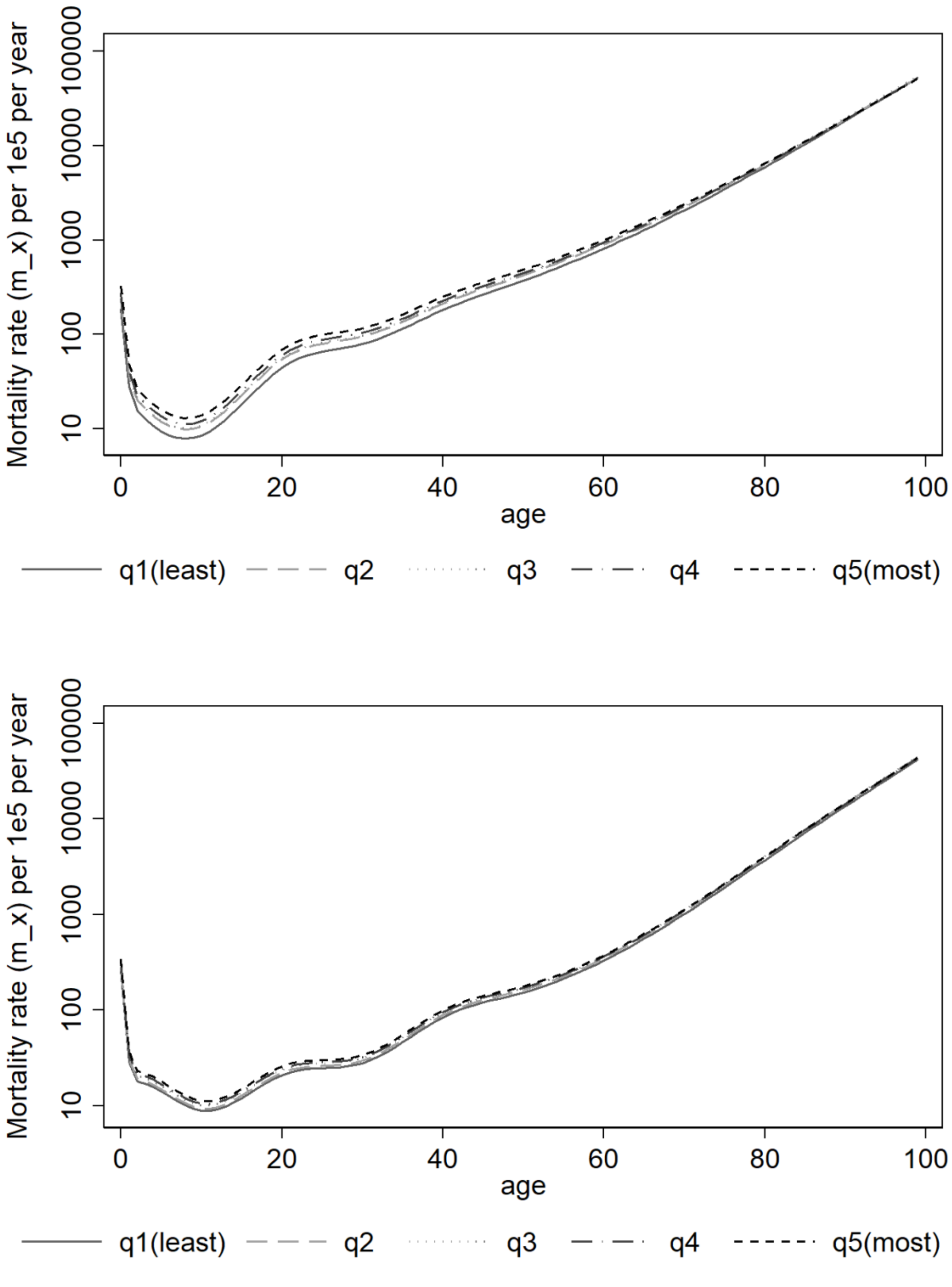

Figure 2 – Predicted mortality rates (log scale) according to quintiles of socioeconomic deprivation 2010-2012 for Men (top) and Women (bottom).

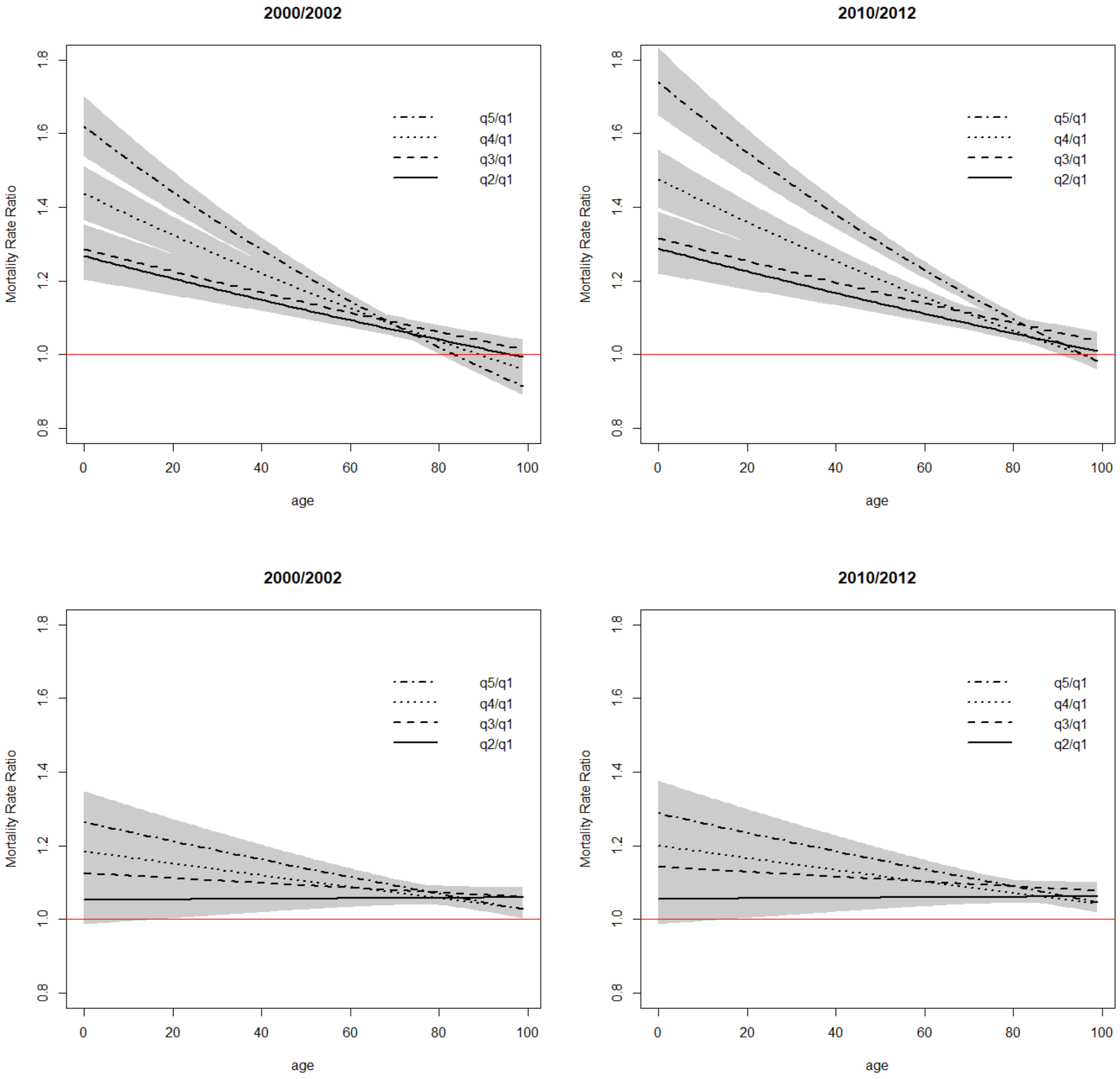

Figure 3 – Mortality Rate Ratio as function of age between deprivation quintiles q2, q3, q4 and q5 and least deprived quintile (q1) for Men (top) and Women (bottom).

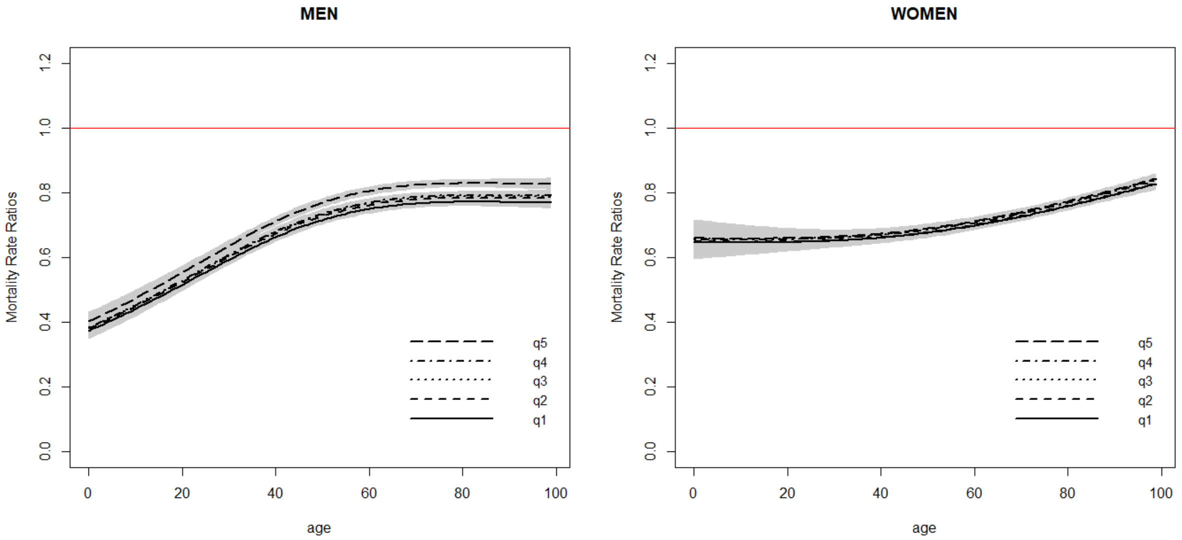

Figure 4 – Mortality Rate Ratio as function of age between period 2010-2012 and 2000-2002 (reference) according to deprivation quintile for men (left) and women (right).

Deprivation-specific life tables using multivariables flexible modelling - trends from 2000-2002 to 2010-2012, Portugal

Luis Antunes, Denisa Mendonça, Ana Isabel Ribeiro, Camille Maringe, Bernard Rachet

Supplementary Material S1

Estimation of confidence intervals for model-based predicted mortality rate ratios

Considering a GLM model with Poisson error for the number of deaths by age, deprivation quintile and period:

$$log(d_{age,i,j}) = \beta_0 + f(age) + \sum_{k=2}^{5} \beta_{d_k} \cdot dep_k + g_1(age * dep_i) + \beta_{period_j} \cdot period_j + \\ + \sum_{k=2}^{5} \beta_{dep\_period_k} \cdot period_j * dep_k + g_2(age * period_j) + log(pyrs_{age,i,j})$$

Reference categories:

Deprivation: first quintile (least deprived)

Period: 2000-02

Considering linear interaction between age and deprivation:

$$g_1(age * dep_i) = \sum_{i=2}^{5} \beta_{age\_dep_i} \cdot age * dep_i$$

Considering interaction of period with splines for age with 1 internal knot:

$$g_2(age * period_j) = \beta_{period\_age_1} \cdot (period_j * age)_S + \beta_{period\_age_2} \cdot v(period_j * age),$$

where $(period_j * age)_S$ represents the standardised $period * age$ variable and $v(period_j * age)$ represents the orthogonalised spline basis.

Mortality rate:

$$log(R_{age,i,j}) = \beta_0 + f(age) + \sum_{k=2}^{5} \beta_{d_k} \cdot dep_k + g_1(age * dep_i) + \beta_{period_j} \cdot period_j + \\ + \sum_{k=2}^{5} \beta_{dep\_period_k} \cdot period_j * dep_i + g_2(age * period_j)$$

where

$$log(R_{age,i,j}) = log(d_{age,i,j}) - log(pyrs_{age,i,j}) = log\left(\frac{d_{age,i,j}}{pyrs_{age,i,j}}\right)$$

Mortality rate ratio between two different deprivation groups, same period:

$$log\left(\frac{R_{age,i=a,j}}{R_{age,i=b,j}}\right) =$$

$$= \beta_0 + f(age) + \beta_{d_{i=a}} + \beta_{age\_dep=a} \cdot age + \beta_{period_j} \cdot period_j + \beta_{dep\_period=a} \cdot period_j +$$

$$+ g_2(age * period_j) -$$

$$[\beta_0 + f(age) + \beta_{d_{i=b}} + \beta_{age\_dep=b} \cdot age + \beta_{period_j} \cdot period_j + \beta_{dep\_period=b} \cdot period_j +$$

$$+ g_2(age * period_j)] =$$

$$= \beta_{d_{i=a}} + \beta_{age\_dep=a} \cdot age + \beta_{dep\_period=a} \cdot period_j - (\beta_{d_{i=b}} + \beta_{age\_dep=b} \cdot age + \beta_{dep\_period=b} \cdot period_j)$$

$$= (\beta_{d_{i=a}} - \beta_{d_{i=b}}) + (\beta_{age\_dep=a} - \beta_{age\_dep=b}) \cdot age + (\beta_{dep\_period=a} - \beta_{dep\_period=b}) \cdot period_j$$

Calculate Confidence Interval assuming normality of $log(RR)$.

The variance of the $log(RR)$ can be estimated using the delta method:

$$VAR[log(RR)(\beta)] \simeq \left[\frac{\partial log(RR)}{\partial \beta}\right]_{\beta=\hat{\beta}} \times VAR[\hat{\beta}] \times \left[\frac{\partial log(RR)}{\partial \beta}\right]^T_{\beta=\hat{\beta}}$$

$$\left[\frac{\partial log(RR)}{\partial \beta}\right]_{\beta=\hat{\beta}} = \begin{bmatrix} 1 & -1 & age & -age & period & -period \end{bmatrix}$$

$$VAR[\hat{\beta}] = \begin{bmatrix} VAR[\beta_{d_{i=a}}] & COV[\beta_{d_{i=a}}, \beta_{d_{i=b}}] & \cdot & \cdot & \cdot & \cdot \\ \cdot & VAR[\beta_{d_{i=b}}] & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & COV[\beta_{dep\_period=b}, \beta_{dep\_period=a}] & VAR[\beta_{dep\_period=b}] \end{bmatrix}$$

The 95% confidence interval for the mortality rate ratio is then given by:

$$exp\left[log\left(\frac{R_{age,i=a,j}}{R_{age,i=b,j}}\right) \pm 1.96 \times \sqrt{VAR[log(RR)]}\right]$$

Considering that the deprivation group $i = b$ is the reference group, the expressions above simplify to:

$$log\left(\frac{R_{age,i=a,j}}{R_{age,i=b,j}}\right) =$$
$$= \beta_{d_{i=a}} + \beta_{age\_dep=a} \cdot age + \beta_{dep\_period=a} \cdot period_j$$

$$\left[\frac{\partial log(RR)}{\partial \beta}\right]_{\beta=\hat{\beta}} = \begin{bmatrix} 1 & age & period \end{bmatrix}$$

$$VAR[\hat{\beta}] = \begin{bmatrix} VAR[\beta_{d_{i=a}}] & COV[\beta_{d_{i=a}}, \beta_{age\_dep=a}] & \cdot \\ \cdot & VAR[\beta_{age\_dep=a}] & \cdot \\ \cdot & COV[\beta_{age\_dep=a}, \beta_{dep\_period=a}] & VAR[\beta_{dep\_period=a}] \end{bmatrix}$$

Mortality rate ratio between two different periods ($period = 1/period = 0$), same deprivation group:

$$log\left(\frac{R_{age,i,j=1}}{R_{age,i,j=0}}\right) =$$
$$= \beta_0 + f(age) + \beta_{d_i} + \beta_{age\_dep=i} + \beta_{period_1} + \beta_{dep\_period=i1} + g_2(age * period_1) -$$
$$- (\beta_0 + f(age) + \beta_{d_i} + \beta_{age\_dep=i} + g_2(age * period_0)) =$$
$$= \beta_{period_1} + \beta_{dep\_period=i1} + g_2(age * period_1) - g_2(age * period_0)$$

Assuming only one knot for the interaction age*period as stated above:

$$log\left(\frac{R_{age,i,j=1}}{R_{age,i,j=0}}\right) = \beta_{period_1} + \beta_{dep\_period=i1} + \beta_{period\_age_1} \cdot ((period_1 * age)_S - (period_0 * age)_S) +$$
$$+ \beta_{period\_age_2} \cdot (v(period_1 * age) - v(period_0 * age))$$

Again, the variance of the $log(RR)$ can be estimated using the delta method:

$$VAR[log(RR)(\beta)] \simeq \left[\frac{\partial log(RR)}{\partial \beta}\right]_{\beta=\hat{\beta}} \times VAR[\hat{\beta}] \times \left[\frac{\partial log(RR)}{\partial \beta}\right]^T_{\beta=\hat{\beta}}$$

$$\left[\frac{\partial log(RR)}{\partial \beta}\right]_{\beta=\hat{\beta}} = \begin{bmatrix} 1 & 1 & (period_1 * age)_S - (period_0 * age)_S & v(period_1 * age) - v(period_0 * age) \end{bmatrix}$$

$$VAR[\hat{\beta}] = \begin{bmatrix} VAR[\beta_{period_1}] & COV[\beta_{period_1}, \beta_{dep\_period=i1}] & . & . \\ . & VAR[\beta_{dep\_period=i1}] & . & . \\ . & . & . & . \\ . & . & COV[\beta_{period\_age_2}, \beta_{period\_age_1}] & VAR[\beta_{period\_age_2}] \end{bmatrix}$$

The 95% confidence interval for the mortality rate ratio is then given by:

$$exp\left[log\left(\frac{R_{age,i,j=1}}{R_{age,i,j=0}}\right) \pm 1.96 \times \sqrt{VAR[log(RR)]}\right]$$

Table S1 - Life tables by deprivation quintile (1-Least deprived) for men in the period 2000-2002 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
0	496,1	75,1	628,8	74,0	638,1	73,8	712,8	73,6	802,9	73,3
1	72,8	74,5	92,0	73,5	93,4	73,3	104,2	73,1	117,1	72,9
2	40,5	73,5	51,0	72,5	51,8	72,3	57,7	72,2	64,7	72,0
3	32,3	72,6	40,6	71,6	41,2	71,4	45,8	71,2	51,4	71,0
4	26,7	71,6	33,5	70,6	34,0	70,4	37,8	70,3	42,3	70,1
5	23,0	70,6	28,8	69,6	29,2	69,4	32,3	69,3	36,1	69,1
6	20,5	69,6	25,6	68,7	26,0	68,5	28,7	68,3	32,0	68,1
7	19,0	68,6	23,7	67,7	24,0	67,5	26,5	67,3	29,5	67,1
8	18,3	67,7	22,8	66,7	23,1	66,5	25,5	66,4	28,3	66,2
9	18,4	66,7	22,8	65,7	23,1	65,5	25,5	65,4	28,2	65,2
10	19,2	65,7	23,7	64,7	24,1	64,5	26,5	64,4	29,3	64,2
11	20,8	64,7	25,7	63,7	26,1	63,5	28,6	63,4	31,6	63,2
12	23,4	63,7	28,8	62,8	29,2	62,5	32,0	62,4	35,3	62,2
13	27,0	62,7	33,1	61,8	33,6	61,6	36,8	61,5	40,5	61,3
14	31,7	61,7	38,9	60,8	39,5	60,6	43,1	60,5	47,4	60,3
15	37,8	60,8	46,2	59,8	46,9	59,6	51,1	59,5	56,1	59,3
16	45,3	59,8	55,2	58,8	56,1	58,6	61,0	58,5	66,8	58,4
17	54,2	58,8	65,9	57,9	66,9	57,7	72,7	57,6	79,5	57,4
18	64,3	57,8	78,0	56,9	79,2	56,7	85,9	56,6	93,8	56,4
19	75,2	56,9	91,0	56,0	92,4	55,8	100,0	55,7	109,1	55,5
20	86,0	55,9	103,7	55,0	105,4	54,8	113,9	54,7	124,0	54,5
21	95,6	55,0	115,1	54,1	117,0	53,9	126,2	53,8	137,1	53,6
22	103,8	54,0	124,6	53,1	126,6	52,9	136,4	52,8	147,9	52,7
23	110,3	53,1	132,1	52,2	134,3	52,0	144,4	51,9	156,3	51,8
24	115,3	52,1	137,8	51,3	140,0	51,1	150,3	51,0	162,5	50,8
25	119,0	51,2	141,9	50,3	144,2	50,1	154,5	50,1	166,8	49,9
26	121,9	50,3	144,9	49,4	147,3	49,2	157,6	49,1	169,8	49,0
27	124,2	49,3	147,4	48,5	149,8	48,3	160,0	48,2	172,1	48,1
28	126,7	48,4	149,9	47,5	152,4	47,3	162,5	47,3	174,4	47,2
29	129,7	47,4	153,1	46,6	155,6	46,4	165,7	46,4	177,6	46,3
30	133,9	46,5	157,7	45,7	160,3	45,5	170,4	45,4	182,3	45,3
31	139,9	45,6	164,3	44,8	167,0	44,6	177,2	44,5	189,3	44,4
32	147,7	44,6	173,0	43,8	175,9	43,6	186,3	43,6	198,7	43,5
33	157,3	43,7	183,9	42,9	187,0	42,7	197,7	42,7	210,5	42,6
34	168,9	42,8	196,9	42,0	200,2	41,8	211,3	41,8	224,6	41,7
35	182,2	41,8	212,0	41,1	215,5	40,9	227,2	40,9	241,0	40,8
36	197,5	40,9	229,1	40,1	233,0	40,0	245,1	39,9	259,6	39,9
37	214,4	40,0	248,2	39,2	252,4	39,1	265,1	39,0	280,3	39,0
38	232,9	39,1	268,9	38,3	273,5	38,2	286,8	38,1	302,7	38,1
39	252,7	38,2	291,0	37,4	296,0	37,3	309,9	37,2	326,5	37,2
40	273,3	37,2	314,0	36,5	319,4	36,4	333,8	36,4	351,1	36,3
41	294,3	36,4	337,3	35,7	343,1	35,5	358,0	35,5	376,0	35,4
42	315,7	35,5	360,9	34,8	367,1	34,6	382,4	34,6	400,9	34,6
43	337,5	34,6	384,9	33,9	391,5	33,7	407,2	33,7	426,1	33,7
44	359,8	33,7	409,3	33,0	416,4	32,9	432,3	32,9	451,7	32,8
45	382,8	32,8	434,5	32,2	442,0	32,0	458,1	32,0	477,8	32,0
46	406,8	31,9	460,5	31,3	468,6	31,1	484,9	31,2	504,8	31,1
47	432,0	31,1	487,9	30,4	496,4	30,3	512,9	30,3	533,1	30,3
48	458,9	30,2	517,0	29,6	526,0	29,4	542,5	29,5	562,9	29,5
49	487,8	29,3	548,2	28,7	557,8	28,6	574,4	28,6	595,0	28,6

Table S1 (cont.) - Life tables by deprivation quintile (1-Least deprived) for men in the period 2000-2002 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
50	519,3	28,5	582,2	27,9	592,4	27,7	609,0	27,8	629,7	27,8
51	554,0	27,6	619,5	27,1	630,4	26,9	647,0	26,9	667,9	27,0
52	592,2	26,8	660,7	26,2	672,4	26,1	688,9	26,1	709,9	26,1
53	634,4	25,9	706,0	25,4	718,6	25,2	735,0	25,3	756,2	25,3
54	681,1	25,1	756,1	24,6	769,6	24,4	785,8	24,5	807,1	24,5
55	732,6	24,2	811,3	23,8	825,8	23,6	841,9	23,7	863,2	23,7
56	789,6	23,4	872,3	22,9	887,9	22,8	903,7	22,9	925,0	22,9
57	852,7	22,6	939,6	22,1	956,6	22,0	971,9	22,1	993,1	22,1
58	922,6	21,8	1014,1	21,3	1032,4	21,2	1047,3	21,3	1068,3	21,3
59	1000,0	21,0	1096,5	20,6	1116,4	20,4	1130,5	20,5	1151,2	20,6
60	1085,8	20,2	1187,7	19,8	1209,3	19,6	1222,6	19,7	1242,9	19,8
61	1181,1	19,4	1288,8	19,0	1312,3	18,9	1324,5	19,0	1344,2	19,0
62	1287,0	18,6	1400,9	18,2	1426,5	18,1	1437,4	18,2	1456,3	18,3
63	1404,7	17,9	1525,3	17,5	1553,3	17,4	1562,5	17,5	1580,3	17,5
64	1535,7	17,1	1663,4	16,8	1694,0	16,6	1701,3	16,7	1717,7	16,8
65	1681,5	16,4	1816,9	16,0	1850,4	15,9	1855,3	16,0	1870,0	16,1
66	1844,1	15,6	1987,6	15,3	2024,4	15,2	2026,3	15,3	2038,9	15,4
67	2025,3	14,9	2177,6	14,6	2218,1	14,5	2216,5	14,6	2226,5	14,7
68	2227,7	14,2	2389,3	13,9	2433,9	13,8	2428,1	13,9	2434,8	14,0
69	2453,7	13,5	2625,3	13,2	2674,4	13,1	2663,6	13,2	2666,4	13,3
70	2706,4	12,8	2888,5	12,6	2942,7	12,5	2925,9	12,6	2924,0	12,7
71	2989,0	12,2	3182,3	11,9	3242,2	11,8	3218,4	12,0	3210,8	12,1
72	3305,4	11,5	3510,5	11,3	3576,8	11,2	3544,6	11,3	3530,2	11,4
73	3659,8	10,9	3877,4	10,7	3950,8	10,6	3908,8	10,7	3886,2	10,8
74	4057,1	10,3	4287,8	10,1	4369,2	10,0	4315,5	10,1	4283,3	10,2
75	4502,8	9,7	4747,1	9,5	4837,5	9,4	4770,1	9,5	4726,4	9,7
76	5002,9	9,1	5261,4	9,0	5361,9	8,9	5278,4	9,0	5221,2	9,1
77	5564,6	8,6	5837,7	8,4	5949,5	8,3	5847,2	8,5	5773,9	8,6
78	6195,6	8,0	6483,8	7,9	6608,4	7,8	6483,9	7,9	6391,6	8,0
79	6905,0	7,5	7208,5	7,4	7347,4	7,3	7197,0	7,4	7082,5	7,5
80	7703,0	7,0	8021,7	6,9	8176,8	6,8	7996,1	6,9	7855,4	7,1
81	8600,8	6,5	8934,7	6,4	9107,9	6,4	8891,8	6,5	8720,5	6,6
82	9611,5	6,1	9960,2	6,0	10153,8	5,9	9896,4	6,0	9689,2	6,2
83	10749,7	5,6	11112,4	5,6	11329,0	5,5	11023,5	5,6	10774,2	5,7
84	12032,0	5,2	12407,4	5,2	12650,0	5,1	12288,4	5,2	11990,0	5,3
85	13477,0	4,8	13863,4	4,8	14135,3	4,7	13708,4	4,8	13352,6	4,9
86	15105,9	4,5	15500,8	4,4	15805,7	4,3	15302,9	4,5	14880,3	4,6
87	16942,6	4,1	17342,8	4,1	17685,0	4,0	17093,9	4,1	16593,4	4,2
88	19014,0	3,8	19415,3	3,7	19799,5	3,7	19105,9	3,8	18514,9	3,9
89	21348,6	3,5	21745,7	3,4	22177,2	3,4	21364,8	3,5	20668,5	3,6
90	23972,1	3,2	24358,1	3,1	24842,8	3,1	23892,9	3,2	23074,8	3,3
91	26918,0	2,9	27284,3	2,9	27828,8	2,8	26720,3	2,9	25761,3	3,0
92	30226,0	2,6	30562,1	2,6	31173,7	2,6	29882,2	2,7	28760,5	2,7
93	33940,4	2,4	34233,6	2,4	34920,7	2,3	33418,2	2,4	32108,9	2,5
94	38111,3	2,2	38346,1	2,1	39118,0	2,1	37372,7	2,2	35847,2	2,3
95	42794,8	1,9	42952,8	1,9	43819,8	1,9	41795,2	2,0	40020,6	2,0
96	48053,8	1,7	48112,8	1,7	49086,8	1,7	46741,0	1,7	44679,9	1,8
97	53959,1	1,4	53892,8	1,4	54986,8	1,4	52272,0	1,4	49881,7	1,5
98	60590,1	1,0	60367,1	1,0	61596,0	1,0	58457,5	1,1	55689,1	1,1
99	68036,0	0,5	67619,1	0,5	68999,6	0,5	65375,0	0,5	62172,7	0,5

Table S2 - Life tables by deprivation quintile (1-Least deprived) for women in the period 2000-2002 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
0	407,7	81,6	429,3	81,0	458,9	80,8	482,9	80,8	515,3	80,6
1	44,1	80,9	46,4	80,4	49,6	80,2	52,1	80,2	55,6	80,0
2	27,8	79,9	29,3	79,4	31,3	79,2	32,9	79,3	35,0	79,1
3	26,5	78,9	27,9	78,4	29,7	78,2	31,2	78,3	33,2	78,1
4	24,3	78,0	25,6	77,5	27,3	77,3	28,7	77,3	30,5	77,1
5	21,9	77,0	23,0	76,5	24,6	76,3	25,7	76,3	27,4	76,2
6	19,5	76,0	20,5	75,5	21,8	75,3	22,9	75,3	24,3	75,2
7	17,3	75,0	18,3	74,5	19,4	74,3	20,3	74,4	21,6	74,2
8	15,6	74,0	16,5	73,5	17,5	73,3	18,3	73,4	19,4	73,2
9	14,4	73,0	15,2	72,5	16,2	72,3	16,9	72,4	17,9	72,2
10	13,8	72,0	14,5	71,5	15,4	71,4	16,1	71,4	17,1	71,2
11	13,7	71,1	14,5	70,6	15,4	70,4	16,0	70,4	17,0	70,2
12	14,2	70,1	15,0	69,6	15,9	69,4	16,6	69,4	17,6	69,3
13	15,2	69,1	16,1	68,6	17,0	68,4	17,7	68,4	18,7	68,3
14	16,7	68,1	17,6	67,6	18,6	67,4	19,4	67,5	20,5	67,3
15	18,6	67,1	19,6	66,6	20,8	66,4	21,6	66,5	22,8	66,3
16	21,0	66,1	22,1	65,6	23,4	65,4	24,3	65,5	25,6	65,3
17	23,7	65,1	25,0	64,6	26,4	64,4	27,4	64,5	28,9	64,3
18	26,7	64,1	28,1	63,6	29,7	63,5	30,8	63,5	32,5	63,3
19	29,7	63,2	31,3	62,7	33,0	62,5	34,2	62,5	36,1	62,4
20	32,4	62,2	34,2	61,7	36,1	61,5	37,3	61,6	39,3	61,4
21	34,6	61,2	36,5	60,7	38,4	60,5	39,8	60,6	41,8	60,4
22	36,1	60,2	38,1	59,7	40,1	59,5	41,5	59,6	43,6	59,4
23	37,1	59,2	39,2	58,7	41,2	58,6	42,6	58,6	44,7	58,5
24	37,7	58,3	39,8	57,8	41,9	57,6	43,2	57,6	45,4	57,5
25	38,1	57,3	40,2	56,8	42,3	56,6	43,6	56,7	45,7	56,5
26	38,4	56,3	40,5	55,8	42,6	55,6	43,9	55,7	46,0	55,5
27	38,9	55,3	41,0	54,8	43,0	54,7	44,3	54,7	46,4	54,6
28	39,6	54,3	41,8	53,9	43,8	53,7	45,1	53,7	47,2	53,6
29	40,9	53,4	43,1	52,9	45,2	52,7	46,5	52,8	48,6	52,6
30	42,9	52,4	45,3	51,9	47,5	51,7	48,7	51,8	50,9	51,6
31	46,0	51,4	48,6	50,9	50,9	50,8	52,2	50,8	54,5	50,7
32	50,3	50,4	53,0	50,0	55,5	49,8	56,9	49,8	59,4	49,7
33	55,7	49,5	58,8	49,0	61,5	48,8	63,0	48,9	65,7	48,7
34	62,5	48,5	65,9	48,0	68,9	47,8	70,5	47,9	73,5	47,8
35	70,6	47,5	74,5	47,0	77,8	46,9	79,5	46,9	82,9	46,8
36	80,0	46,6	84,4	46,1	88,1	45,9	90,0	46,0	93,8	45,8
37	90,7	45,6	95,7	45,1	99,9	44,9	101,9	45,0	106,1	44,9
38	102,5	44,6	108,2	44,2	112,8	44,0	115,0	44,1	119,6	43,9
39	115,0	43,7	121,4	43,2	126,4	43,0	128,8	43,1	133,9	43,0
40	127,6	42,7	134,7	42,3	140,2	42,1	142,8	42,2	148,3	42,0
41	139,6	41,8	147,4	41,3	153,4	41,2	156,1	41,2	162,0	41,1
42	150,9	40,8	159,4	40,4	165,7	40,2	168,4	40,3	174,7	40,2
43	161,4	39,9	170,4	39,4	177,1	39,3	179,9	39,4	186,5	39,2
44	171,2	39,0	180,7	38,5	187,7	38,4	190,5	38,4	197,3	38,3
45	180,3	38,0	190,4	37,6	197,5	37,4	200,3	37,5	207,4	37,4
46	189,0	37,1	199,6	36,6	207,0	36,5	209,7	36,6	216,9	36,4
47	197,6	36,2	208,7	35,7	216,2	35,6	218,9	35,6	226,3	35,5
48	206,4	35,2	218,0	34,8	225,8	34,6	228,4	34,7	236,0	34,6
49	215,9	34,3	228,1	33,9	236,0	33,7	238,6	33,8	246,3	33,7

Table S2 (cont.) - Life tables by deprivation quintile (1-Least deprived) for women in the period 2000-2002 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
50	226,6	33,4	239,4	32,9	247,6	32,8	250,0	32,9	257,9	32,8
51	239,0	32,5	252,4	32,0	260,9	31,9	263,3	32,0	271,4	31,9
52	253,2	31,5	267,5	31,1	276,3	31,0	278,6	31,0	287,1	30,9
53	269,7	30,6	285,0	30,2	294,1	30,0	296,3	30,1	305,1	30,0
54	288,6	29,7	304,9	29,3	314,5	29,1	316,6	29,2	325,7	29,1
55	310,2	28,8	327,7	28,3	337,8	28,2	339,8	28,3	349,4	28,2
56	334,8	27,9	353,8	27,4	364,5	27,3	366,3	27,4	376,4	27,3
57	362,9	27,0	383,5	26,5	394,8	26,4	396,4	26,5	407,1	26,4
58	394,9	26,1	417,4	25,6	429,4	25,5	430,8	25,6	442,1	25,5
59	431,4	25,2	456,0	24,7	468,8	24,6	470,0	24,7	482,0	24,6
60	473,0	24,3	500,0	23,9	513,7	23,7	514,6	23,8	527,3	23,7
61	520,4	23,4	550,2	23,0	564,9	22,9	565,3	23,0	579,0	22,9
62	574,5	22,5	607,3	22,1	623,1	22,0	623,1	22,1	637,7	22,0
63	636,1	21,6	672,5	21,2	689,6	21,1	689,0	21,2	704,7	21,1
64	706,4	20,8	746,9	20,4	765,3	20,3	764,1	20,4	780,9	20,3
65	786,6	19,9	831,8	19,5	851,8	19,4	849,7	19,5	867,8	19,4
66	878,3	19,0	928,7	18,7	950,4	18,6	947,3	18,7	966,9	18,6
67	983,0	18,2	1039,5	17,8	1063,1	17,7	1058,7	17,8	1079,9	17,8
68	1102,6	17,4	1166,1	17,0	1191,8	16,9	1185,9	17,0	1208,8	17,0
69	1239,4	16,6	1310,8	16,2	1338,8	16,1	1331,1	16,2	1355,8	16,2
70	1395,7	15,8	1476,3	15,4	1506,8	15,3	1496,9	15,4	1523,7	15,4
71	1574,5	15,0	1665,5	14,6	1698,8	14,6	1686,2	14,7	1715,3	14,6
72	1779,0	14,2	1881,9	13,9	1918,2	13,8	1902,5	13,9	1934,0	13,8
73	2012,7	13,5	2129,3	13,1	2169,0	13,1	2149,4	13,2	2183,5	13,1
74	2279,9	12,7	2412,2	12,4	2455,5	12,3	2431,3	12,4	2468,3	12,4
75	2585,3	12,0	2735,4	11,7	2782,7	11,6	2753,0	11,7	2793,0	11,7
76	2934,0	11,3	3104,6	11,0	3156,2	10,9	3120,0	11,0	3163,1	11,0
77	3332,0	10,6	3526,0	10,3	3582,2	10,3	3538,2	10,4	3584,7	10,3
78	3786,0	10,0	4006,6	9,7	4067,9	9,6	4014,5	9,7	4064,6	9,7
79	4303,2	9,3	4554,3	9,1	4620,8	9,0	4556,5	9,1	4610,2	9,1
80	4891,8	8,7	5177,6	8,5	5249,8	8,4	5172,5	8,5	5229,9	8,5
81	5560,8	8,1	5886,2	7,9	5964,3	7,9	5871,6	7,9	5932,8	7,9
82	6320,2	7,6	6690,4	7,3	6774,7	7,3	6663,9	7,4	6728,9	7,4
83	7180,6	7,0	7601,7	6,8	7692,4	6,8	7560,4	6,9	7629,0	6,8
84	8153,8	6,5	8632,5	6,3	8729,7	6,3	8572,9	6,4	8644,8	6,3
85	9252,3	6,0	9796,2	5,8	9899,9	5,8	9714,0	5,9	9789,0	5,9
86	10489,5	5,6	11106,8	5,4	11217,0	5,4	10997,3	5,4	11074,8	5,4
87	11879,5	5,1	12579,4	5,0	12695,9	4,9	12437,0	5,0	12516,1	5,0
88	13437,1	4,7	14229,8	4,6	14352,0	4,5	14047,8	4,6	14127,7	4,6
89	15177,6	4,3	16074,0	4,2	16201,4	4,2	15844,8	4,2	15924,3	4,2
90	17121,4	4,0	18133,8	3,8	18265,5	3,8	17848,8	3,9	17926,2	3,9
91	19309,1	3,6	20452,2	3,5	20587,1	3,5	20100,9	3,5	20174,6	3,5
92	21776,3	3,3	23067,0	3,2	23203,9	3,1	22637,2	3,2	22704,9	3,2
93	24558,8	3,0	26016,1	2,8	26153,2	2,8	25493,5	2,9	25552,5	2,9
94	27696,8	2,6	29342,3	2,5	29477,4	2,5	28710,2	2,6	28757,4	2,6
95	31235,8	2,3	33093,7	2,2	33224,1	2,2	32332,8	2,3	32364,1	2,3
96	35227,0	2,0	37324,7	1,9	37447,1	1,9	36412,4	1,9	36423,3	1,9
97	39728,1	1,6	42096,7	1,6	42206,8	1,6	41006,9	1,6	40991,5	1,6
98	44804,4	1,1	47478,8	1,1	47571,6	1,1	46181,0	1,1	46132,7	1,1
99	50529,3	0,5	53548,9	0,5	53618,2	0,5	52008,0	0,5	51918,7	0,5

Table S3 - Life tables by deprivation quintile (1-Least deprived) for men in the period 2010-2012 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
0	185,4	78,8	238,7	77,8	243,9	77,5	273,5	77,4	322,3	76,7
1	27,7	77,9	35,5	76,9	36,3	76,7	40,6	76,6	47,8	76,0
2	15,6	76,9	20,0	76,0	20,5	75,7	22,9	75,6	26,9	75,0
3	12,7	75,9	16,2	75,0	16,6	74,7	18,5	74,6	21,7	74,0
4	10,7	75,0	13,6	74,0	13,9	73,7	15,5	73,6	18,1	73,0
5	9,3	74,0	11,9	73,0	12,1	72,7	13,5	72,7	15,8	72,1
6	8,5	73,0	10,7	72,0	11,0	71,7	12,2	71,7	14,2	71,1
7	8,0	72,0	10,1	71,0	10,3	70,8	11,4	70,7	13,3	70,1
8	7,8	71,0	9,9	70,0	10,1	69,8	11,2	69,7	13,0	69,1
9	8,0	70,0	10,0	69,0	10,3	68,8	11,3	68,7	13,2	68,1
10	8,5	69,0	10,6	68,0	10,9	67,8	12,0	67,7	13,9	67,1
11	9,3	68,0	11,7	67,1	12,0	66,8	13,2	66,7	15,2	66,1
12	10,7	67,0	13,3	66,1	13,6	65,8	15,0	65,7	17,3	65,1
13	12,5	66,0	15,6	65,1	15,9	64,8	17,5	64,7	20,1	64,1
14	14,9	65,0	18,6	64,1	19,0	63,8	20,8	63,7	23,9	63,1
15	18,1	64,0	22,4	63,1	22,9	62,8	25,1	62,8	28,8	62,2
16	22,0	63,0	27,2	62,1	27,8	61,8	30,4	61,8	34,8	61,2
17	26,7	62,1	33,0	61,1	33,7	60,9	36,8	60,8	42,1	60,2
18	32,2	61,1	39,7	60,1	40,6	59,9	44,1	59,8	50,4	59,2
19	38,2	60,1	46,9	59,2	48,0	58,9	52,2	58,8	59,5	58,3
20	44,3	59,1	54,4	58,2	55,6	57,9	60,3	57,9	68,7	57,3
21	50,1	58,1	61,2	57,2	62,6	57,0	67,8	56,9	77,1	56,3
22	55,1	57,2	67,3	56,3	68,8	56,0	74,4	55,9	84,4	55,4
23	59,4	56,2	72,3	55,3	74,0	55,0	79,9	55,0	90,5	54,4
24	63,0	55,2	76,5	54,3	78,3	54,1	84,3	54,0	95,4	53,5
25	66,0	54,3	79,9	53,4	81,8	53,1	87,9	53,1	99,3	52,5
26	68,5	53,3	82,7	52,4	84,7	52,2	90,9	52,1	102,5	51,6
27	70,8	52,3	85,3	51,5	87,3	51,2	93,6	51,2	105,3	50,6
28	73,1	51,4	87,9	50,5	90,0	50,2	96,3	50,2	108,2	49,7
29	75,9	50,4	91,0	49,5	93,1	49,3	99,5	49,3	111,6	48,7
30	79,3	49,5	94,9	48,6	97,1	48,3	103,6	48,3	116,0	47,8
31	83,9	48,5	100,1	47,6	102,5	47,4	109,2	47,4	122,0	46,8
32	89,7	47,5	106,8	46,7	109,3	46,4	116,2	46,4	129,6	45,9
33	96,7	46,6	114,8	45,7	117,6	45,5	124,8	45,5	139,0	45,0
34	105,0	45,6	124,4	44,8	127,3	44,5	134,9	44,5	150,0	44,0
35	114,6	44,7	135,5	43,8	138,7	43,6	146,7	43,6	162,9	43,1
36	125,6	43,7	148,0	42,9	151,6	42,7	160,1	42,6	177,4	42,2
37	137,8	42,8	162,1	42,0	166,0	41,7	175,0	41,7	193,6	41,2
38	151,3	41,8	177,5	41,0	181,7	40,8	191,3	40,8	211,3	40,3
39	165,8	40,9	194,0	40,1	198,7	39,9	208,8	39,9	230,2	39,4
40	181,1	40,0	211,4	39,2	216,5	38,9	227,1	38,9	250,0	38,5
41	196,8	39,0	229,2	38,3	234,8	38,0	245,9	38,0	270,2	37,6
42	213,0	38,1	247,5	37,3	253,5	37,1	265,0	37,1	290,7	36,7
43	229,7	37,2	266,2	36,4	272,7	36,2	284,6	36,2	311,6	35,8
44	246,9	36,3	285,4	35,5	292,4	35,3	304,7	35,3	333,1	34,9
45	264,8	35,4	305,4	34,6	312,8	34,4	325,5	34,4	355,2	34,0
46	283,6	34,5	326,2	33,7	334,1	33,5	347,1	33,5	378,1	33,1
47	303,3	33,5	348,0	32,8	356,6	32,6	369,8	32,6	402,1	32,2
48	324,4	32,6	371,3	32,0	380,4	31,7	393,8	31,8	427,5	31,4
49	347,1	31,8	396,3	31,1	406,0	30,8	419,7	30,9	454,8	30,5

Table S3 (cont.) - Life tables by deprivation quintile (1-Least deprived) for men in the period 2010-2012 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
50	371,8	30,9	423,5	30,2	433,9	30,0	447,7	30,0	484,4	29,6
51	398,9	30,0	453,3	29,3	464,5	29,1	478,5	29,2	516,8	28,8
52	428,9	29,1	486,1	28,5	498,1	28,2	512,3	28,3	552,3	27,9
53	461,9	28,2	522,2	27,6	535,2	27,4	549,5	27,4	591,4	27,1
54	498,3	27,3	562,0	26,7	576,0	26,5	590,4	26,6	634,4	26,2
55	538,5	26,5	605,9	25,9	621,0	25,7	635,5	25,7	681,7	25,4
56	583,0	25,6	654,3	25,0	670,7	24,8	685,1	24,9	733,7	24,6
57	632,1	24,8	707,8	24,2	725,5	24,0	739,9	24,1	791,0	23,8
58	686,6	23,9	766,8	23,4	786,1	23,2	800,4	23,2	854,2	22,9
59	746,9	23,1	832,1	22,5	853,1	22,3	867,1	22,4	923,8	22,1
60	813,8	22,2	904,4	21,7	927,2	21,5	941,0	21,6	1000,8	21,3
61	888,0	21,4	984,5	20,9	1009,4	20,7	1022,6	20,8	1085,7	20,5
62	970,4	20,6	1073,2	20,1	1100,4	19,9	1113,0	20,0	1179,7	19,8
63	1062,0	19,8	1171,6	19,3	1201,4	19,1	1213,1	19,2	1283,6	19,0
64	1163,9	19,0	1280,9	18,6	1313,5	18,4	1324,1	18,5	1398,7	18,2
65	1277,3	18,2	1402,3	17,8	1438,1	17,6	1447,2	17,7	1526,1	17,5
66	1403,7	17,5	1537,2	17,0	1576,6	16,8	1584,0	17,0	1667,5	16,7
67	1544,6	16,7	1687,4	16,3	1730,7	16,1	1735,9	16,2	1824,3	16,0
68	1701,8	16,0	1854,6	15,6	1902,3	15,4	1904,9	15,5	1998,4	15,3
69	1877,4	15,2	2040,9	14,8	2093,5	14,7	2092,8	14,8	2191,9	14,6
70	2073,6	14,5	2248,6	14,1	2306,7	14,0	2302,1	14,1	2407,0	13,9
71	2292,9	13,8	2480,4	13,4	2544,6	13,3	2535,3	13,4	2646,3	13,2
72	2538,4	13,1	2739,1	12,8	2810,2	12,6	2795,3	12,7	2912,7	12,6
73	2813,2	12,4	3028,2	12,1	3107,0	12,0	3085,4	12,1	3209,4	11,9
74	3121,1	11,8	3351,4	11,5	3438,7	11,3	3409,2	11,4	3540,2	11,3
75	3466,3	11,1	3712,9	10,8	3809,9	10,7	3770,9	10,8	3909,1	10,7
76	3853,5	10,5	4117,6	10,2	4225,4	10,1	4175,2	10,2	4320,8	10,1
77	4288,1	9,9	4570,7	9,6	4690,6	9,5	4627,2	9,6	4780,4	9,5
78	4776,2	9,3	5078,5	9,1	5212,0	8,9	5132,9	9,1	5293,8	9,0
79	5324,5	8,7	5647,6	8,5	5796,4	8,4	5699,1	8,5	5867,6	8,4
80	5941,0	8,2	6286,0	8,0	6452,0	7,9	6333,1	8,0	6509,2	7,9
81	6634,2	7,6	7002,3	7,5	7187,6	7,3	7043,4	7,5	7227,0	7,4
82	7414,3	7,1	7806,4	7,0	8013,4	6,9	7839,6	7,0	8030,2	6,9
83	8292,3	6,6	8709,5	6,5	8940,9	6,4	8732,4	6,5	8929,4	6,5
84	9281,0	6,2	9724,0	6,0	9982,9	5,9	9733,9	6,1	9936,5	6,0
85	10394,7	5,7	10864,1	5,6	11154,0	5,5	10857,8	5,6	11064,8	5,6
86	11649,6	5,3	12145,8	5,2	12470,7	5,1	12119,2	5,2	12329,2	5,2
87	13064,1	4,9	13587,1	4,8	13951,3	4,7	13535,6	4,8	13746,7	4,8
88	14658,9	4,5	15208,3	4,4	15616,9	4,3	15126,4	4,5	15336,0	4,4
89	16456,0	4,1	17030,9	4,1	17489,4	4,0	16911,9	4,1	17116,9	4,1
90	18475,1	3,8	19073,5	3,7	19588,1	3,7	18909,8	3,8	19106,4	3,7
91	20741,9	3,5	21361,2	3,4	21938,8	3,4	21143,9	3,4	21327,2	3,4
92	23286,8	3,2	23923,3	3,1	24571,5	3,1	23641,8	3,1	23806,1	3,1
93	26144,0	2,8	26792,6	2,8	27520,1	2,8	26434,9	2,8	26573,1	2,8
94	29351,8	2,6	30006,1	2,5	30822,6	2,5	29557,9	2,5	29661,7	2,5
95	32953,1	2,2	33605,1	2,2	34521,4	2,2	33049,9	2,3	33109,3	2,3
96	36996,3	1,9	37635,6	1,9	38664,0	1,9	36954,4	1,9	36957,6	1,9
97	41535,6	1,6	42149,7	1,6	43303,8	1,5	41320,3	1,6	41253,2	1,6
98	46631,9	1,1	47205,1	1,1	48500,4	1,1	46201,9	1,1	46048,1	1,1
99	52353,4	0,5	52866,8	0,5	54320,6	0,5	51660,2	0,5	51400,4	0,5

Table S4 - Life tables by deprivation quintile (1-Least deprived) for women in the period 2010-2012 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
0	263,6	84,3	278,3	83,8	301,2	83,6	316,3	83,6	339,4	83,4
1	28,5	83,6	30,1	83,1	32,5	82,8	34,1	82,9	36,6	82,7
2	18,0	82,6	19,0	82,1	20,5	81,8	21,5	81,9	23,1	81,7
3	17,1	81,6	18,1	81,1	19,5	80,8	20,4	80,9	21,9	80,7
4	15,7	80,6	16,6	80,1	17,9	79,9	18,8	79,9	20,1	79,8
5	14,1	79,6	14,9	79,1	16,1	78,9	16,8	79,0	18,0	78,8
6	12,6	78,6	13,3	78,2	14,3	77,9	15,0	78,0	16,0	77,8
7	11,2	77,7	11,8	77,2	12,8	76,9	13,3	77,0	14,2	76,8
8	10,1	76,7	10,7	76,2	11,5	75,9	12,0	76,0	12,8	75,8
9	9,3	75,7	9,8	75,2	10,6	74,9	11,0	75,0	11,8	74,8
10	8,9	74,7	9,4	74,2	10,1	73,9	10,5	74,0	11,2	73,8
11	8,9	73,7	9,4	73,2	10,1	72,9	10,5	73,0	11,2	72,8
12	9,2	72,7	9,7	72,2	10,4	71,9	10,9	72,0	11,6	71,8
13	9,8	71,7	10,4	71,2	11,2	70,9	11,6	71,0	12,3	70,8
14	10,8	70,7	11,4	70,2	12,2	70,0	12,7	70,0	13,5	69,9
15	12,0	69,7	12,7	69,2	13,6	69,0	14,1	69,1	15,0	68,9
16	13,6	68,7	14,3	68,2	15,4	68,0	15,9	68,1	16,9	67,9
17	15,3	67,7	16,2	67,2	17,3	67,0	18,0	67,1	19,0	66,9
18	17,2	66,7	18,2	66,3	19,5	66,0	20,2	66,1	21,4	65,9
19	19,2	65,8	20,3	65,3	21,7	65,0	22,4	65,1	23,8	64,9
20	21,0	64,8	22,2	64,3	23,7	64,0	24,5	64,1	25,9	63,9
21	22,4	63,8	23,7	63,3	25,3	63,0	26,1	63,1	27,6	62,9
22	23,4	62,8	24,7	62,3	26,4	62,1	27,2	62,1	28,8	62,0
23	24,1	61,8	25,4	61,3	27,1	61,1	27,9	61,2	29,5	61,0
24	24,5	60,8	25,9	60,3	27,6	60,1	28,4	60,2	30,0	60,0
25	24,7	59,8	26,1	59,4	27,8	59,1	28,6	59,2	30,2	59,0
26	24,9	58,9	26,4	58,4	28,1	58,1	28,8	58,2	30,4	58,0
27	25,2	57,9	26,7	57,4	28,4	57,1	29,1	57,2	30,7	57,1
28	25,7	56,9	27,2	56,4	28,9	56,2	29,7	56,2	31,3	56,1
29	26,6	55,9	28,1	55,4	29,9	55,2	30,6	55,3	32,2	55,1
30	28,0	54,9	29,6	54,4	31,4	54,2	32,1	54,3	33,8	54,1
31	30,0	53,9	31,7	53,4	33,7	53,2	34,4	53,3	36,2	53,1
32	32,8	52,9	34,7	52,5	36,8	52,2	37,6	52,3	39,5	52,1
33	36,4	52,0	38,5	51,5	40,8	51,2	41,7	51,3	43,8	51,2
34	40,9	51,0	43,2	50,5	45,8	50,3	46,7	50,4	49,0	50,2
35	46,2	50,0	48,9	49,5	51,7	49,3	52,8	49,4	55,3	49,2
36	52,5	49,0	55,5	48,6	58,7	48,3	59,8	48,4	62,7	48,2
37	59,6	48,0	63,1	47,6	66,6	47,3	67,8	47,4	71,0	47,3
38	67,4	47,1	71,4	46,6	75,3	46,4	76,7	46,5	80,2	46,3
39	75,8	46,1	80,2	45,6	84,6	45,4	86,0	45,5	89,9	45,3
40	84,2	45,1	89,2	44,7	94,0	44,4	95,5	44,5	99,8	44,4
41	92,4	44,2	97,8	43,7	103,0	43,5	104,6	43,6	109,2	43,4
42	100,1	43,2	105,9	42,8	111,5	42,5	113,1	42,6	118,0	42,5
43	107,2	42,3	113,5	41,8	119,4	41,6	121,0	41,7	126,2	41,5
44	114,0	41,3	120,6	40,8	126,9	40,6	128,4	40,7	133,8	40,6
45	120,3	40,4	127,4	39,9	133,8	39,7	135,4	39,8	141,0	39,6
46	126,4	39,4	133,9	38,9	140,6	38,7	142,1	38,8	147,9	38,7
47	132,5	38,4	140,3	38,0	147,2	37,8	148,7	37,9	154,7	37,7
48	138,8	37,5	147,0	37,1	154,1	36,8	155,6	36,9	161,7	36,8
49	145,6	36,6	154,2	36,1	161,6	35,9	162,9	36,0	169,2	35,9

Table S4 (cont.) - Life tables by deprivation quintile (1-Least deprived) for women in the period 2010-2012 (m_x - mortality rate; e_x - life expectancy at age x).

age	EDI = 1		EDI = 2		EDI = 3		EDI = 4		EDI = 5	
	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x	m_x	e_x
50	153,2	35,6	162,2	35,2	169,9	34,9	171,2	35,1	177,7	34,9
51	162,0	34,7	171,6	34,2	179,6	34,0	180,8	34,1	187,5	34,0
52	172,2	33,7	182,4	33,3	190,8	33,1	191,9	33,2	198,9	33,0
53	184,0	32,8	194,9	32,3	203,7	32,1	204,7	32,2	212,0	32,1
54	197,5	31,8	209,2	31,4	218,5	31,2	219,4	31,3	227,1	31,2
55	212,9	30,9	225,6	30,5	235,5	30,3	236,3	30,4	244,4	30,2
56	230,6	30,0	244,3	29,5	254,9	29,3	255,5	29,4	264,1	29,3
57	250,8	29,0	265,7	28,6	277,0	28,4	277,5	28,5	286,7	28,4
58	273,9	28,1	290,2	27,7	302,3	27,5	302,6	27,6	312,4	27,5
59	300,2	27,2	318,1	26,8	331,3	26,6	331,3	26,7	341,7	26,5
60	330,3	26,3	350,1	25,8	364,3	25,6	364,0	25,8	375,2	25,6
61	364,8	25,3	386,6	24,9	402,0	24,7	401,4	24,9	413,5	24,7
62	404,1	24,4	428,4	24,0	445,1	23,8	444,1	24,0	457,2	23,8
63	449,2	23,5	476,1	23,1	494,4	22,9	492,9	23,1	507,1	22,9
64	500,8	22,6	530,8	22,2	550,9	22,1	548,7	22,2	564,1	22,1
65	559,8	21,7	593,5	21,3	615,5	21,2	612,5	21,3	629,3	21,2
66	627,5	20,9	665,3	20,5	689,5	20,3	685,6	20,4	703,9	20,3
67	705,1	20,0	747,6	19,6	774,3	19,4	769,3	19,6	789,3	19,4
68	794,1	19,1	842,1	18,7	871,5	18,6	865,2	18,7	887,1	18,6
69	896,3	18,3	950,5	17,9	983,1	17,7	975,1	17,9	999,1	17,8
70	1013,6	17,4	1074,9	17,1	1111,0	16,9	1101,1	17,0	1127,5	16,9
71	1148,2	16,6	1217,7	16,2	1257,9	16,1	1245,7	16,2	1274,6	16,1
72	1302,8	15,8	1381,8	15,4	1426,4	15,3	1411,4	15,4	1443,2	15,3
73	1480,3	15,0	1570,2	14,6	1619,8	14,5	1601,4	14,6	1636,5	14,5
74	1684,1	14,2	1786,4	13,9	1841,7	13,7	1819,3	13,8	1857,8	13,8
75	1918,0	13,4	2034,6	13,1	2096,2	13,0	2069,0	13,1	2111,4	13,0
76	2186,2	12,7	2319,4	12,4	2388,0	12,2	2355,0	12,4	2401,7	12,3
77	2493,8	12,0	2645,9	11,6	2722,3	11,5	2682,5	11,6	2733,9	11,6
78	2846,2	11,3	3019,9	10,9	3105,1	10,8	3057,2	10,9	3113,6	10,9
79	3249,5	10,6	3448,1	10,3	3543,0	10,2	3485,4	10,3	3547,4	10,2
80	3710,6	9,9	3937,6	9,6	4043,3	9,5	3974,4	9,6	4042,4	9,6
81	4237,2	9,2	4496,7	9,0	4614,4	8,9	4532,0	9,0	4606,4	8,9
82	4837,7	8,6	5134,3	8,4	5265,2	8,3	5166,9	8,4	5248,2	8,3
83	5521,4	8,0	5860,3	7,8	6005,7	7,7	5888,8	7,8	5977,4	7,7
84	6298,4	7,5	6685,5	7,2	6846,8	7,1	6708,0	7,2	6804,4	7,2
85	7179,7	6,9	7621,5	6,7	7800,3	6,6	7635,8	6,7	7740,3	6,7
86	8177,2	6,4	8680,9	6,2	8878,7	6,1	8684,4	6,2	8797,3	6,2
87	9303,5	5,9	9877,3	5,7	10095,6	5,6	9866,5	5,7	9988,1	5,7
88	10572,0	5,4	11224,7	5,2	11465,3	5,2	11195,9	5,3	11326,2	5,2
89	11996,5	5,0	12738,1	4,8	13002,5	4,7	12686,5	4,8	12825,6	4,8
90	13595,6	4,5	14436,9	4,4	14726,8	4,3	14357,1	4,4	14504,8	4,4
91	15403,7	4,1	16358,0	4,0	16675,5	3,9	16243,4	4,0	16399,5	4,0
92	17452,2	3,7	18534,7	3,6	18882,0	3,6	18377,6	3,6	18541,7	3,6
93	19773,2	3,3	21001,0	3,2	21380,4	3,2	20792,2	3,2	20963,7	3,2
94	22402,9	3,0	23795,6	2,9	24209,4	2,8	23524,0	2,9	23702,1	2,9
95	25382,3	2,6	26962,0	2,5	27412,8	2,5	26614,7	2,5	26798,2	2,5
96	28758,0	2,2	30549,7	2,1	31040,0	2,1	30111,5	2,1	30298,8	2,1
97	32582,6	1,7	34614,9	1,7	35147,2	1,7	34067,7	1,7	34256,6	1,7
98	36915,8	1,2	39220,9	1,2	39797,8	1,2	38543,8	1,2	38731,4	1,2
99	41825,3	0,5	44439,9	0,5	45063,8	0,5	43607,9	0,5	43790,8	0,5

3.4 Study IV: Methods to deal with missing data in excess hazard model covariates

Dealing with missing information on covariates of excess hazard models - making the imputation model compatible with the substantive model

Luís Antunes, Denisa Mendonça, Maria José Bento, Edmund-Njeru Njagi, Aurélien Belot, Bernard Rachet

A large proportion of cases in the population-based cancer datasets have missing information on stage of disease at diagnosis. Ignoring the cases with missing information can lead to biased results and conclusions especially if the mechanism of missingness is not completely at random. Among the different ways of dealing with missing data, multiple imputation has become more available in common statistical software packages and is increasingly used. However careful should be given to its proper use. The incompatibility between the imputation and substantive model, which can arise when the associations between variables in the substantive model are not taken into account in the imputation models or when the substantive model is itself nonlinear, can lead to invalid inference. Motivated by the analysis of population-based cancer survival analysis, the multiple imputation substantive model compatible fully conditional specification (SMC-FCS) approach, proposed by Bartlett and colleagues in 2015, was extended in this study to accommodate excess hazard models. The proposed approach was compared with the standard fully conditional (FCS) multiple imputation procedure and with the complete-case analysis (CCA) using a simulation study. The SMC-FCS approach produced unbiased estimates in all scenarios tested, while the standard FCS produced biased estimates and poor empirical coverages probabilities. CCA only produced biased estimates for missingness mechanism dependent on the outcome.

The three approaches were then used in a study which aimed at evaluating socioeconomic inequalities in survival from cancer (more specifically for a cohort of colorectal cancer patients diagnosed in the North region of Portugal). The socioeconomic effects were adjusted for age, sex and extent of disease at diagnosis. This last covariate had missing information for around 40% of the cases. Deprivation-specific life tables were used to adjust for background mortality.

Although statistically significant differences in crude net survival were observed between socioeconomic groups, after adjusting for the extent of disease using an excess hazard model, the inequalities diminished and were no longer significant. These conclusions were transversal to the three approaches used.

Although the distribution of imputed values was similar between the two MI approaches, the effect of extent of disease on the excess hazard seemed to be diluted in the SMC-FCS approach. Further research is warranted to analyse the performance of this approach when imputing categorical variables.

Next, the resulting manuscript of this study is presented.

Dealing with missing information on covariates of excess hazard models – making the imputation model compatible with the substantive model

Luís Antunes[1,2,3], Denisa Mendonça[3,4], Maria José Bento[1], Edmund-Njeru Njagi[5], Aurélien Belot[5], Bernard Rachet[5]

[1] Grupo de Epidemiologia de Cancro, Centro de Investigação do IPO Porto (CI-IPOP), Instituto Português de Oncologia do Porto (IPO Porto), Porto, Portugal

[2] Faculdade de Ciências, Universidade do Porto, Portugal

[3] EPI-UNIT - Instituto de Saúde Pública, Universidade do Porto, Porto, Portugal

[4] Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Portugal

[5] Cancer Survival Group, London School of Hygiene and Tropical Medicine, United Kingdom

Corresponding author:

Luis Antunes

Rua Dr. António Bernardino de Almeida

4200-072 Porto, Portugal

Email: luis.antunes@ipoporto.min-saude.pt

Abstract

Missing data is a common issue in epidemiological databases. Ignoring the cases with missing information can lead to biased results and conclusions especially if the mechanism of missingness is not completely at random. Among the different ways of dealing with missing data, multiple imputation has become more available in common statistical software packages and is increasingly used. However attention should be given to its proper use. The incompatibility between the imputation and substantive model, which can arise when the associations between variables in the substantive model are not taken into account in the imputation models or when the model is itself nonlinear, can lead to invalid inference.

Aiming at analysing population-based cancer survival data, we extended the multiple imputation substantive model compatible fully conditional specification (SMC-FCS) approach, proposed by Bartlett and colleagues in 2015, to accommodate excess hazard models. The proposed approach was compared with the standard fully conditional (FCS) multiple imputation procedure and with the complete-case analysis (CCA) using a simulation study. The SMC-FCS approach produced unbiased estimates in all scenarios tested, while the standard FCS produced biased estimates and poor empirical coverages probabilities. The SMC-FCS algorithm was then used in the evaluation of socioeconomic inequalities in survival from cancer. A cohort of colorectal cancer patients diagnosed in the North Region of Portugal was analysed. No major differences were observed in the estimated effects of deprivation level between the three approaches analysed. In none of the scenarios were observed significant adjusted effects of the deprivation level. The SMC-FCS tended to bias effect of extent of disease towards the null. Further research is warranted to better evaluate the performance of the SMC-FCS algorithm in the imputation of categorical variables and to extend it to cope with time-dependent effects.

Keywords: Missing data, multiple imputation, substantive model compatible, excess hazard, socioeconomic inequalities

1. Introduction

Missing data is an almost unavoidable issue in observational studies. Due to multiple possible reasons, incomplete information on the outcomes or in the covariates is likely to occur. Multiple imputation (MI) has become in the last years one of the most common methodologies for handling missing data [1,2]. Its increasing availability in common statistical packages made the application of MI more attractive to a larger spectrum of users. This broader application of the methodology was not necessarily followed by a correct application or reporting of the same. Rezvan and colleagues systematically reviewed manuscripts published during six years in two important medical journals in which multiple imputation was carried out [2]. From the 103 articles identified, only 37% described the imputation model, only two compared the imputed with the observed values and only three performed sensitivity analysis.

Also, the problem of incompatibility between imputation model and the substantive (or analysis) model can lead to invalid inference. This problem can occur when the substantive model includes nonlinear covariate effects, interactions or when the model itself is nonlinear (e.g. hazard models).

When the outcome of interest is survival time and there is missing information on covariates, it is consensual that the outcome should be included in the imputation model. However, different ways of including the survival outcome can be found in the literature: the censoring indicator (δ) and the survival time (*T*) [3]; δ and *log(T)* [4,5]; δ, *log(T)* and *T* [6]. In 2009, White and Royston [7] recommended the inclusion of the cumulative baseline hazard ($\Lambda_0(t)$) besides the censor indicator in the imputation model when the substantive model of interest is a Cox hazard model and showed that the result is exact in the case of a single binary covariate and in other cases approximately valid for small covariate effects and/or small cumulative incidence. In 2015, Bartlett and colleagues developed an algorithm for MI that ensures the compatibility between both models and designated it as Substantive Model Compatible Fully Conditional Specification (SMC-FCS) [8]. This method has been implemented in STATA and R but only a limited number of substantive models are available [9]. Recently, Keogh and Morris [10] extended this approach to hazard models with time-varying effects of covariates.

In population-based cancer survival analysis, the interest normally lies on the excess hazard modelling. Excess hazard represents the hazard due to the disease and is now commonly modelled using flexible parametric models [11,12]. In this relative survival framework, multiple imputation has also been used to deal with missing information on excess hazard model covariates [13–18]. In 2015, Falcaro and colleagues evaluated the use of MI in the context of net survival problems with missing information, more specifically, in the excess hazard modelling using flexible parametric proportional hazards models with missing data on categorical covariates (stage of disease at diagnosis) [19]. The results obtained suggested that a multinomial logistic imputation model for stage should be used and that the Nelson-Aalen cumulative hazard estimate and the event indicator should be included in the imputation models, as already suggested by White and Royston in the context of the Cox model. The issue of compatibility between the imputation and substantive models when these are excess hazard models has however still not been properly addressed.

The main aim of this work was to extend the SMC-FCS algorithm developed by Bartlett and colleagues to accommodate excess hazard models. The performance of the extension proposed was compared with the standard fully conditional specification (FCS) approach and with a complete-case analysis (CCA), using a simulation study. The three methods were then applied to a survival dataset from a cohort of colorectal cancer patients extracted from the North Region of Portugal Cancer Registry (RORENO).

The article is organised as follows. In Section 2, an overview of the methods used in this study is given and the proposed extension of the SMC-FCS algorithm for excess hazard models is presented. A simulation study evaluating the performance of the SMC-FCS algorithm is presented in Section 3. The motivating dataset is characterised in Section 4 and then analysed in Section 5 with the aim of evaluating socioeconomic inequalities in survival from cancer when adjusting for extent of disease at diagnosis. The article finishes with a discussion in Section 6.

2. Methods

2.1 Excess hazard modelling

In population-based cancer survival analysis, since cause of death is usually unknown or unreliable, the analysis is performed in the relative survival setting. It is considered that the observed hazard (λ_O) can be decomposed in two additive parcels, the cancer related hazard (excess hazard) (λ_E) and the other causes hazard (λ_P), estimated by the general population mortality: $\lambda_O = \lambda_P + \lambda_E$. The excess hazard function is modelled as a function of a set of covariates. A flexible parametric model for the excess hazard function is considered here:

$$\lambda_E(t, \boldsymbol{X}) = \lambda_0(t) \cdot \exp\big(g(\boldsymbol{X})\big),$$

where λ_0(t) is the excess hazard baseline. Following the formulation of Charvat and colleagues [20], the baseline was modelled using B-spline functions. Covariate effects can be considered linear or non-linear and time-dependent effects can also be easily added in this formulation.

2.2 Multiple imputation

Multiple imputation (MI) was first introduced by Rubin in 1978 [21]. Initially, MI was developed in the framework of survey nonresponse but has nowadays been expanded to broader set of different fields, including survival analysis [22].

In MI several imputations are generated for each missing value, as opposed to single imputation where each missing value is replaced by a single value. This creates several completed datasets, as many as the number of imputations performed. Each completed dataset is analysed using standard methods for complete data. The results from the several analyses are then combined to produce single estimates and confidence intervals that incorporate missing-data uncertainty.

The process can be divided in three main steps: the imputation, the analysis and the combination steps. The models related to the first step are commonly designated as imputation models and the ones used in the second step, as substantive models [23]. Briefly the algorithm goes like this:

i. Using the imputation model, generate *M>1* values for each missing value, obtaining *M* completed datasets;
ii. Fit the substantive model independently to each one of the *M* completed datasets;
iii. Combine the results obtained from each analysis performed in the previous step using Rubin's rules [24].

The MI algorithm typically relies on the assumption that the data are missing at random (MAR). This means that the probability of having a missing observation is random conditioned on the observed information, i.e. does not depend on unobserved data.

In MI the imputation phase is separated from the analysis phase. The imputation models used may thus be incompatible with the substantive model. This means that there is no joint model for which the conditionals equal the imputation and substantive conditional models.

2.3 Compatibility between imputation and substantive model

To overcome the problem of incompatibility between imputation and substantive models in multiple imputation, Bartlett and colleagues [8] developed an algorithm that ensures that each covariate with missing observations is imputed from a model compatible with the substantive model. The algorithm is referred as Substantive Model Compatible-Fully Conditional Specification (SMC-FCS).

The rational of the method is described briefly. Let *Y* represent the outcome, ***X*** a vector of *p* partially observed covariates and ***Z*** a vector of fully observed covariates. For each partially observed covariate X_j, $\boldsymbol{X}_{-j}$ represents the vector of covariates excluding that covariate (X_1, …,X_{j-1}, X_{j+1}, …, X_p). Bartlett starts by noting that the imputation model for X_j, conditioned on the remaining covariates and the outcome is proportional to the product of the substantive model and the imputation model for X_j not involving the outcome:

$$f\left(X_j \middle| X_{-j}, Z, Y\right) = \frac{f\left(Y, X_j, X_{-j}, Z\right)}{f\left(Y, X_{-j}, Z\right)}$$

$$\propto f(Y|X,Z) \cdot f(X_j|X_{-j}, Z)$$

So, in the algorithm SMC-FCS, a model $f(X_j|X_{-j},Z,\phi_j)$ must be specified for each $j=1,\dots,p$, together with noninformative priors $f(\phi_j)$. Given values of the parameters of the imputation and substantive model (ϕ_j and ψ, respectively) the missing values of X_j are imputed from a density proportional to:

$$f(Y|X,Z,\psi) \cdot f(X_j|X_{-j}, Z, \phi_j)$$

Since generally this density does not belong to a standard parametric family, drawing samples from it is non-trivial [8]. Bartlett and colleagues proposed a rejection sampling procedure that involves repeatedly drawing samples from a candidate distribution, *$f(X_j|X_{-j},Z,\phi_j)$*, until the drawn value X_j satisfies the condition:

$$U \leq \frac{f(Y|X_j^*, X_{-j}, Z, \psi)}{c(Y, X_{-j}, Z, \psi)}$$

where *U* follows an uniform distribution on (0,1) and *$c(Y, X_{-j}, Z, \psi)$* is an upper bound (in X_j) for *$f(Y|Xj,X\text{-}j,Z,\psi)$* that does not involve X_j.

2.4 SMC-FCS in excess hazard models

The SMC-FCS algorithm was extended here to accommodate excess hazard models. A detailed description on the derivation of the conditions in which the rejection sampling must be done is presented in the Supplementary Material (Section S1).

We consider that the substantive model of interest is an excess hazard model with *p* partially observed variables $\boldsymbol{X} = (X_1, \ldots, X_p)$ and a fully observed vector of variables $\boldsymbol{Z}=(Z_1, \ldots, Z_q)$:

$$\lambda_E(\boldsymbol{X}, \boldsymbol{Z}, t; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_0(t; \boldsymbol{\gamma}) \cdot \exp(g(\boldsymbol{X}, \boldsymbol{Z}))$$

The algorithm to generate the m^{th} imputed data set is as follows (adapted from [10]):

1) Calculate using the population mortality, the population hazard (λ_P) and the cumulative population hazard (Λ_P) given the demographic variables. This does not depend on the imputed values so it must be done only once.

2) Fill in all missing values for the incomplete variables with a starting arbitrary value (for example, mean or mode of observed values).

3) Fit the excess hazard model of interest to the current complete dataset to obtain estimates of the model parameters $(\hat{\beta}, \hat{\gamma})$ and of the respective variance-covariance matrix $\hat{\Sigma}$. Draw values $\beta^{(m)}$ and $\gamma^{(m)}$ from a joint normal distributions with means $\hat{\beta}$ and $\hat{\gamma}$ and variance-covariance matrices $\hat{\Sigma}$.

4) Calculate the estimate of the baseline excess hazard $\lambda_0^{(m)}(t)$ and of the baseline cumulative excess hazard $\Lambda_0^{(m)}(t)$ using parameter values $\gamma^{(m)}$.

5) Fit a regression model (linear, logistic, multinomial, as appropriate) of X_j on X_{-j} and Z to the current completed data set. Draw a value ϕ^* from the approximate joint posterior distribution of ϕ.

6) For each individual for whom X_j is missing, (i) draw a value of X_j^* from the distribution $f(X_j|X_{-j},Z;\phi^*)$ and, (ii) draw a value U from a uniform distribution on [0,1]. Accept the value X^* if:

$$U \leq exp[-\Lambda_P(t)] \cdot exp\left[-\Lambda_0^{(m)}(t) \cdot e^{g(X_j^*,X_{-j},Z,\beta)}\right] \qquad \text{for } \delta = 0$$

$$U \leq \frac{\left[\lambda_P(t) + \lambda_0^{(m)}(t) \cdot e^{g(X_j^*,X_{-j},Z,\beta)}\right] \cdot exp\left[-\Lambda_P(t) - \Lambda_0^{(m)}(t) \cdot e^{g(X_j^*,X_{-j},Z,\beta)}\right]}{\frac{\lambda_0(t)}{\Lambda_0(t)} \cdot exp\left[-\Lambda_P(t) - 1 + \frac{\Lambda_0(t) \cdot \lambda_P(t)}{\lambda_0(t)}\right]}$$

for δ = 1

Repeat (i) and (ii) until a value of X_j^* is accepted.

7) Return to step 3 until one cycle is done for all variables with missing data.

8) Repeat steps 3-7 a certain number of iterations so that the imputed values of **X** convergence to a stationary distribution. The obtained values form the *m*th imputed data set. Repeat the process *M* times to obtain *M* imputed datasets.

3. Simulation study

A simulation study was first performed to evaluate the performance of the SMC-FCS algorithm when the substantive model of interest is an excess hazard model. This example was adapted from the one presented by Bartlett and colleagues for the Cox model [8]. Two covariates were simulated, one binary variable X_1~Be(p=0.5) and one continuous $X_2|X_1$~N(μ=X_1, σ=1). Times to death from cancer were simulated from the excess hazard model: $\lambda_E(t|X) = 0.002\exp(\beta_1 X_1 + \beta_2 X_2)$ considering $\beta_1 = \beta_2 = 1$. Times to death from other causes were generated from an exponential distribution with hazard 0.001. Censoring times were also generated from an exponential distribution but with hazard 0.002. Each of the 1000 simulated datasets had *n*=1000 subjects. Data on X_2 were made missing considering a MCAR mechanism such that the probability of missingness was 0.3. Missingness in X_1 was imposed considering three different scenarios: A) MCAR with probability of missingness 0.3; B) MAR independent of outcome such that *logit* (P(X_1 miss)) = 0.11 - 0.1X_2; C) MAR dependent on outcome such that *logit* (P(X_1 miss)) = - 0.30 + 0.01T (where *T* represents survival time). In the two last scenarios the coefficients were chosen so that the proportion of missingness in X_1 was also around 0.3.

For each simulated dataset, three approaches for handling missing data were compared: i) Complete-case analysis (CCA), where all the cases with at least one variable missing were discarded; ii) Multiple imputation using fully conditional specification (FCS), including in the imputation models the Nelson-Aalen cumulative hazard estimates, the event indicator and X_1 when imputing X_2 or X_2 when imputing X_1: a logistic regression model was used for imputing X_1 and a linear regression model for X_2; iii) Multiple imputation using the substantive model compatible- fully conditional specification algorithm (SMC-FCS) as described above. Again, a logistic regression model was used for imputing X_1 and a linear regression model for imputing X_2. In this algorithm, the outcomes are not included as covariates in the imputation models.

The results obtained for the three simulated scenarios are presented in Table 1. As expected, the CCA produced unbiased estimates of the two model parameters and empirical coverages close to the nominal level of 95% except when the missingness depended on the outcome (Scenario C). The conventional multiple imputation approach (FCS) produced biased estimates for both parameters and empirical coverages below 95% for all scenarios. On the contrary, the SMC-FCS algorithm produced unbiased estimates in all situations, with lower variability than CCA estimates (lower standard deviations) and with empirical coverages within the expected values.

4. Motivating dataset

Colorectal cancer in the North region of Portugal

The North Region Cancer Registry of Portugal (RORENO) is a population-based cancer registry responsible for collecting information on all incidence cancer cases occurring in the North region of Portugal (~3.6 million inhabitants). The registry was setup in 1988 and in 2018 was integrated in the National Cancer Registry (RON).

A previous study [25] evaluated the existence of socioeconomic inequalities in net survival from colorectal cancer diagnosed in the period 2000-2002 in the area covered by RORENO. In that study, we found inequalities in net survival when using general life tables but that disappeared when inducing relatively small socioeconomic differences in background mortality. In the present study, we intended to update that evaluation for a more recent period, using deprivation-specific life tables recently built [*submitted*] and considering extent of disease at diagnosis as a confounder. Extent of disease is a classification defined by the European Network of Cancer Registries (ENCR) based on the TNM classification [26]. The classification is as follows: Tumour localised (T1-2N0M0); Tumour with local spread (T3-4N0M0); Tumour with regional spread (anyTN+M0); Advanced cancer (anyTanyNM1).

More specifically, all new cancer cases of colorectal cancer (ICD10: C18-C20), diagnosed in the period 2010-2012, in patients with age at diagnosis aged at least 15 years-old and below 95, residing in the North region of Portugal, were considered eligible for analysis. Only the first tumour occurring during the analysed period was considered. Second primary colorectal cancers, either synchronous or metachronous were excluded.

Survival time was considered as time between diagnosis and death from any cause or end of follow-up (31st December 2017).

Deprivation indicator

The Portuguese version of the European Deprivation Index was used as deprivation indicator. This index was built using a methodology first proposed by Pornet and colleagues in 2012[27] and then applied to five European countries: France, England, Italy, Spain and Portugal [28]. The index is based on census variables available for each country that are most associated with variables identified from the European Union Statistics on Income and Living Conditions (EU-SILC) survey [29]. The index for Portugal based on 2001 census includes percentage of: non-owned households, households without indoor flushing, residents with low education level (≤6th grade), household with 5 rooms or less, unemployed looking for a job, female residents aged 65 years or more, households without bath/shower and percentage of residents employed in manual occupations [30]. A score was obtained for each parish based on the census responses of its inhabitants. This score was then categorized in five quintiles from the least deprived (q1) to the most deprived (q5) such that each quintile corresponded to 20% of the Portuguese population. Each deceased was assigned with the deprivation quintile corresponding to his/her parish of residence at the time of death.

Data description

A total of 8108 new cancer cases was considered eligible for analysis. After excluding patients with unknown status at the end of follow-up and zero survival time (n=154; 1.9%), a total of 7954 patients was included in the analysis. Distribution of cases by age group, cancer site, deprivation quintile and extent of disease at diagnosis was calculated by sex (Table 2). Male patients represented 58.6% of the cohort. Women presented a higher median age compared to men: 71 vs 69 years (p<0.001). The proportion of rectum cancer cases was higher in men (p=0.035). No differences were found in the distribution by deprivation groups between male and female patients (p=0.208). Also, the distribution of extent of disease at diagnosis was similar between both sexes (p=0.206).

Missing data

A very low proportion of cases had missing information on deprivation quintile (0.5%) being extent of disease at diagnosis the main prognostic variable with a considerable proportion of missing data (40.4%). No differences were found between male and female patients regarding proportion of missing extent. All other variables analysed were significantly associated with extent missingness, even after adjusting for the effect of the other variables using a multivariable logistic regression model (Table 3). Older, colon cancer and least deprived patients had increased odds of having extent of disease at diagnosis unknown. Patients without microscopically verified diagnosis and patients dying in less than 30 days after diagnosis were also associated with increased odds of missing extent.

Age-standardised net survival (ASNS) at 1-year of the patients with known extent (84.2%; 95%CI: 83.1-85.2) was significantly higher than ASNS of patients with missing extent information (80.7%; 95%CI: 79.4-82.1). On the contrary, ASNS at 5-years was higher in patients with unknown extent (67.1%; 95%CI: 65.2-69.1) than with known extent (63.9%; 95%CI: 62.3-65.6).

5. Socioeconomic inequalities in survival from colorectal cancer

The main aim of the analysis performed was to evaluate the existence of socioeconomic inequalities in net survival from colorectal cancer in the cohort of patients described above. Possible confounder variables considered were age, sex and extent of disease at diagnosis. The proportion of cases with missing extent was around 40%.

First, net survival by SE group was estimated for the full dataset using the non-parametric Pohar-Perme estimator [31]. Differences between net survival curves were assessed using the log-rank-type test developed by Grafféo and colleagues [32].

The unadjusted net survival curves (Figure 1) showed a better net survival for patients living in least deprived areas (p=0.010). Five-year net survival was 66.9% for the least deprived group and 62.0% for the most deprived one.

Second, excess hazard ratios were estimated. Missing data was handled using complete-case analysis and multiple imputation using the standard FCS and the adapted SMC-FCS approach. Covariates considered in the model were age, deprivation index (EDI), sex and extent of disease at diagnosis. All covariates were assumed to have no time-dependent effects. The excess hazard baseline was modelled using B-splines with one knot at one year of follow-up.

In this example, only one covariate had missing data (extent). The imputation model in the standard FCS approach included as covariates age, sex, EDI, tumour site and basis of diagnosis besides the event indicator and the cumulative excess hazard estimated by the Nelson-Aalen estimator. In the SMC-FCS approach, the same variables were used in the imputation model except the outcome, namely the cumulative excess hazard baseline and the event indicator. In both MI approaches, extent of disease was imputed using a multinomial logistic regression model. Fifty imputations were used in each approach.

The distribution of the imputed extent of disease values was investigated and compared with the distribution of extent in the complete cases (Supplementary Material S2). The distribution of the values imputed by the standard FCS approach was very similar to the observed using the SMC-FCS. In both, the proportion of cases in the "Advanced" extent was slightly higher (+3.3%) than the proportion observed for the complete cases. On the contrary, for all other three categories, the proportion of imputed cases was lower than the one in the complete cases. We further investigated the relationship between the imputed values and the survival time (Table S2.2). While for the imputed values using the FCS approach, higher extent consistently resulted in lower mean survival time, for the imputed values using SMC-FCS the association between survival time and extent (for the extents “Local” to “Regional”) seems to have been lost.

The results obtained for the excess hazard ratios (EHR) using the three different approaches are presented in Table 4. The estimated EHRs using the complete-case analysis and the FCS approach were similar. Using SMC-FCS, there was attenuation on the differences in hazard between the several categories and the reference category (“Local”).

Independently of the approach used, no significant effects of socioeconomic factors were observed when adjusting for age, sex and extent of disease at diagnosis.

6. Discussion

The SMC-FCS approach to MI was first proposed by Bartlett and colleagues to ensure the compatibility of the imputation models with the substantive model [8]. The algorithm relies on a rejection sampling scheme. The conditions of acceptance of a proposed imputation value depend on the substantive model of interest. These conditions were derived in this study for the situation where the substantive model is an excess hazard model. This type of models is very common in population-based cancer survival analysis.

The proposed adaptation of the SMC-FCS algorithm to cope with excess hazard models was tested in a simulation study for three different scenarios of missingness. When missingness was MCAR or MAR outcome independent, the complete-case analysis produced unbiased estimates as expected. In the third scenario, where missingness was dependent on the outcome (survival time), the model parameters estimates obtained were biased, including the parameter of the variable for which missing mechanism was MCAR. The standard FCS multiple imputation approach produced biased estimates and poor empirical coverages for both parameters. These results were observed in all the three missingness scenarios analysed. Due to the non-linear nature of the substantive model considered (excess hazard model), the FCS approach does not guarantee the compatibility between the imputation and substantive models. On the contrary, the SMC-FCS approach to MI produced unbiased estimates of both parameters in all scenarios. Also, the standard errors of the estimates were lower than for the complete-case analysis. These results confirm that also when the substantive model is an excess hazard model, the SMC-FCS approach has a higher performance relatively to the other two approaches.

One of the advantages associated with multiple imputation is the possibility of using variables in the imputation model that are not of interest in the substantive model, to increase the plausibility of the MAR assumption and the efficiency of the imputation process. In the SMC-FCS algorithm, to draw imputations that are compatible with the substantive model the variables considered in both imputation and substantive models must be the same. In the analysis step one can however use fewer variables. In the example analysed, two auxiliary variables were used in the imputation step (basis of diagnosis and tumour site) since that have been shown to be related with the chance of extent being missing but were not included in the substantive model.

No major differences in the estimated adjusted effects of socioeconomic condition on the excess hazard were observed between the CCA and both MI approaches. There were however differences between the estimated effects of extent of disease between the SMC-FCS approach and the other two approaches, which suggest an attenuation of the extent effect. Further research is warranted to better evaluate the performance of the SMC-FCS algorithm in the imputation of categorical variables (with more than 2 categories).

In MI the missing values are imputed using imputation models dependent on a set of covariates. The efficiency of these imputations depends if there are variables available that are both associated with the probability of missingness and with the missing variable. In this study, the number of variables used in the imputation model was low and their association with extent of disease was weak which can have diminished the efficiency of the imputations performed.

In this study, the proportional hazards assumption was assumed for all variables. We acknowledge that the effect of some covariates can typically be time-dependent. A first approach for extending the SMC-FCS approach to cope with excess hazard models was presented. Further research must be developed to include time-dependent effects in excess hazard models following the work that Keogh and Morris have done for the Cox models [10].

Competing interests

The authors declare that they have no competing interest.

Ethics approval and consent to participate

This study was approved by the Ethical Committee of the Portuguese Oncology Institute of Porto, Portugal.

References

1. Murray JS. Multiple Imputation: A Review of Practical and Theoretical Findings. Stat Sci. Institute of Mathematical Statistics; 2018;33:142–59.

2. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol. BioMed Central; 2015;15:30.

3. Barzi F, Woodward M. Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies. Am J Epidemiol. Oxford University Press; 2004;160:34–45.

4. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data. J Clin Epidemiol. 2003;56:28–37.

5. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol. 2010;10:7.

6. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681–94.

7. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. Wiley-Blackwell; 2009;28:1982–98.

8. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Stat Methods Med Res. 2015;24.

9. Bartlett JW, Morris TP. Multiple imputation of covariates by substantive-model compatible fully conditional specification. Stata J. College Station, TX: Stata Press; 2015;15:437–456(20).

10. Keogh RH, Morris TP. Multiple imputation in Cox regression when there are time-varying effects of covariates. Stat Med. Wiley-Blackwell; 2018;

11. Uhry Z, Bossard N, Remontet L, Iwaz J, Roche L, GRELL EUROCARE-5 Working Group and the CENSUR Working Survival Group. New insights into survival trend analyses in cancer population-based studies. Eur J Cancer Prev. 2017;26:S9–15.

12. Belot A, Remontet L, Rachet B, Dejardin O, Charvat H, Bara S, et al. Describing the association between socioeconomic inequalities and cancer survival: methodological guidelines and illustration with population-based data. Clin Epidemiol. 2018;10:561–73.

13. Giorgi R, Belot A, Gaudart J, Launoy G, French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. Stat Med. 2008;27:6310–31.

14. Nur U, Shack LG, Rachet B, Carpenter JR, Coleman MP. Modelling relative survival in the presence of incomplete data: a tutorial. Int J Epidemiol. 2010;39:118–28.

15. Dejardin O, Rachet B, Morris E, Bouvier V, Jooste V, Haynes R, et al. Management of colorectal cancer explains differences in 1-year relative survival between France and England for patients diagnosed 1997-2004. Br J Cancer. Nature Publishing Group; 2013;108:775–83.

16. Dejardin O, Jones AP, Rachet B, Morris E, Bouvier V, Jooste V, et al. The influence of geographical access to health care and material deprivation on colorectal cancer survival: Evidence from France and England. Heal Place. Elsevier; 2014;30:36–44.

17. Walters S, Maringe C, Butler J, Rachet B, Barrett-Lee P, Bergh J, et al. Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: A population-based study. Br J Cancer. 2013;108:1195–208.

18. Le Guyader-Peyrou S, Orazio S, Dejardin O, Maynadié M, Troussard X, Monnereau A. Factors related to the relative survival of patients with diffuse large B-cell lymphoma in a population-based study in France: does socio-economic status have a role? Haematologica. Ferrata Storti Foundation; 2017;102:584–92.

19. Falcaro M, Nur U, Rachet B, Carpenter JR. Estimating Excess Hazard Ratios and Net Survival When Covariate Data Are Missing. Epidemiology. 2015;26:421–8.

20. 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. 2016;35:3066–84.

21. Rubin DB. Multiple imputations in sample surveys - A phenomenological Bayesian

approach to nonresponse. Proc Surv Res Merhods Sect rhe Am Stat Assoc. 1978. p. 20–34.

22. Carpenter JR, Kenward MG. Multiple Imputation and its Application. First. Chichester, UK: John Wiley & Sons, Ltd; 2013.

23. Carpenter JR, Kenward MG. Missing data in randomised controlled trials a practical guide. Birmingham: Health Technology Assessment Methodology Programme; 2007.

24. Rubin DB, Wiley InterScience. Multiple imputation for nonresponse in surveys. Wiley; 1987.

25. Antunes L, Mendonça D, Bento MJ, Rachet B. No inequalities in survival from colorectal cancer by education and socioeconomic deprivation - a population-based study in the North Region of Portugal, 2000-2002. BMC Cancer. 2016;16.

26. Berrino F, Brown C, Moller T, Sobin L. ENCR RECOMMENDATIONS, Condensed TNM for Coding the Extent of Disease. Lyon; 2002.

27. Pornet C, Delpierre C, Dejardin O, Grosclaude P, Launay L, Guittet L, et al. Construction of an adaptable European transnational ecological deprivation index: the French version. J Epidemiol Community Health. 2012;66:982–9.

28. Guillaume E, Pornet C, Dejardin O, Launay L, Lillini R, Vercelli M, et al. Development of a cross-cultural deprivation index in five European countries. J Epidemiol Community Health. 2015;jech-2015-205729.

29. Eurostat. Access to Microdata. EUROPEAN UNION STATISTICS ON INCOME AND LIVING CONDITIONS (EU-SILC) [Internet]. 2015 [cited 2018 May 19]. Available from: http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-

and-living-conditions

30. Ribeiro AI, Mayer A, Miranda A, Pina M de F de. Acta Médica Portuguesa. Acta Med. Port. 2017.

31. Perme MP, Stare J, Estève J. On Estimation in Relative Survival. Biometrics. 2012;68:113–20.

32. Grafféo N, Castell F, Belot A, Giorgi R. A log-rank-type test to compare net survival distributions. Biometrics. 2016;72:760–9.

Table 1 – Comparison of excess hazard models parameters estimates for different approaches of missing data handling. Results from n=1000 simulations.

	CCA			FCS			SMC-FCS		
	Mean	SD	Cov	Mean	SD	Cov	Mean	SD	Cov
Scenario A									
$\beta_1 = 1$	1.001	0.143	*95.2*	0.929	0.124	*93.4*	1.003	0.126	*95.6*
$\beta_2 = 1$	1.004	0.069	*95.7*	0.858	0.053	*50.7*	1.004	0.057	*95.8*
Scenario B									
$\beta_1 = 1$	1.002	0.155	*94.0*	0.753	0.122	*55.5*	1.003	0.127	*95.2*
$\beta_2 = 1$	0.999	0.136	*94.7*	0.885	0.055	*69.5*	1.009	0.061	*95.1*
Scenario C									
$\beta_1 = 1$	0.855	0.128	*89.4*	0.895	0.128	*89.5*	1.008	0.128	*95.4*
$\beta_2 = 1$	0.819	0.068	*44.7*	0.880	0.051	*62.5*	1.001	0.058	*95.5*

Scenario A: X1, X2 MCAR

Scenario B: X1 MAR independent of outcome, X2 MCAR

Scenario C: X1 MAR dependent of outcome, X2 MCAR

CCA – Complete-case analysis

FCS – Fully conditional specification

SMC-FCS – Substantive model compatible FCS

Table 2 - Sociodemographic and clinical characteristics of the colorectal cancer patients (2010-2012).

Variable	Male		Female		Total	
	n	%	n	%	n	%
Total by sex	4664	58.6	3290	41.4	7954	100.0
Age group						
15-44	177	3.8	153	4.7	330	4.1
45-54	460	9.9	334	10.2	794	10.0
55-64	1072	23.0	662	20.1	1734	21.8
65-74	1415	30.3	845	25.7	2260	28.4
75+	1540	33.0	1296	39.4	2836	35.7
Tumour site						
Colon	3060	65.6	2234	67.9	5294	66.6
Rectum	1604	34.4	1056	32.1	2660	33.4
Deprivation (EDI)						
q1 (least deprived)	444	9.5	337	10.2	781	9.8
q2	609	13.1	415	12.6	1024	12.9
q3	1074	23.0	693	21.1	1767	22.2
q4	1233	26.4	894	27.2	2127	26.7
q5 (most deprived)	1280	27.4	939	28.5	2219	27.9
unknown	24	0.5	12	0.4	36	0.5
Tumour extent at diagnosis						
Localised	486	10.4	327	9.9	813	10.2
Local spread	782	16.8	510	15.5	1292	16.2
Regional spread	879	18.8	617	18.8	1496	18.8
Advanced	636	13.6	502	15.3	1138	14.3
Unknown	1881	40.3	1334	40.5	3215	40.4

Table 3 - Sociodemographic characteristics of patients with known extent vs patients with unknown extent. Odds ratio of having missing extent (uni and multivariable analysis).

	Extent of disease at diagnosis							
Variable	Known		Unknown		non-adjusted		adjusted	
	n	%	n	%	OR	95%CI	OR	95%CI
Total by extent	4739	59.6	3215	40.4				
Sex								
Male	2783	58.7	1881	58.5	1			
Female	1956	41.3	1334	41.5	1.01	0.92 - 1.11		
Age group								
15-44	225	4.7	105	3.3	1		1	
45-54	512	10.8	282	8.8	1.18	0.90 - 1.55	1.19	0.90 - 1.56
55-64	1088	23.0	646	20.1	1.27	0.99 - 1.64	1.23	0.96 - 1.59
65-74	1362	28.7	898	27.9	1.41	1.10 - 1.81	1.34	1.04 - 1.72
75+	1552	32.7	1284	39.9	1.77	1.39 - 2.26	1.58	1.23 - 2.02
Tumour site								
Colon	2966	62.6	2328	72.4	1		1	
Rectum	1773	37.4	887	27.6	0.64	0.58 - 0.70	0.65	0.59 - 0.72
Deprivation (EDI)								
q1 (least deprived)	430	9.1	351	10.9	1		1	
q2	631	13.3	393	12.2	0.76	0.63 - 0.92	0.72	0.60 - 0.87
q3	1055	22.3	712	22.1	0.83	0.70 - 0.98	0.79	0.66 - 0.93
q4	1258	26.5	869	27.0	0.85	0.72 - 0.99	0.80	0.68 - 0.95
q5 (most deprived)	1351	28.5	868	27.0	0.79	0.67 - 0.93	0.75	0.63 - 0.88
unknown	14	0.3	22	0.7			*a)*	
Basis of diagnosis								
Microscopically verified	4628	97.7	2925	91.0	1		1	
Non-micros. verified	111	2.3	290	9.0	4.13	3.31 - 5.17	3.60	2.86 - 4.53
Death within 30 days of diagnosis								
No	4604	97.2	2994	93.1	1		1	
Yes	135	2.8	221	6.9	2.52	2.02-3.13	1.72	1.36 - 2.16

a) Due to small proportion of cases, this group has been excluded from the logistic regression model.

Table 4 - Excess hazard ratios (CCA; FCS MI; SMC-FCS MI)

Variable	CC		FCS MI		SMC-FCS MI	
	EHR[a)]	95%CI	EHR[a)]	95%CI	EHR[a)]	95%CI
EDI						
q1	1		1		1	
q2	1,01	0,80 - 1,26	1,00	0,82 - 1,22	0,97	0,61 - 1,55
q3	1,03	0,84 - 1,28	1,10	0,93 - 1,31	1,01	0,66 - 1,55
q4	1,11	0,91 - 1,35	1,08	0,91 - 1,28	1,03	0,70 - 1,51
q5	1,09	0,90 - 1,34	1,16	0,98 - 1,37	1,08	0,71 - 1,65
Extent						
Localized	1		1		1	
Local spread	2,69	1,72 - 4,18	2,46	1,61 - 3,75	0,66	0,30 - 1,43
Regional spread	5,41	3,54 - 8,24	5,14	3,52 - 7,53	1,59	0,83 - 3,01
Advanced	33,7	22,3 - 51,1	31,6	21,7 - 46,1	10,5	6,20 - 17,7

a) Adjusted for age, sex and EDI or Extent.

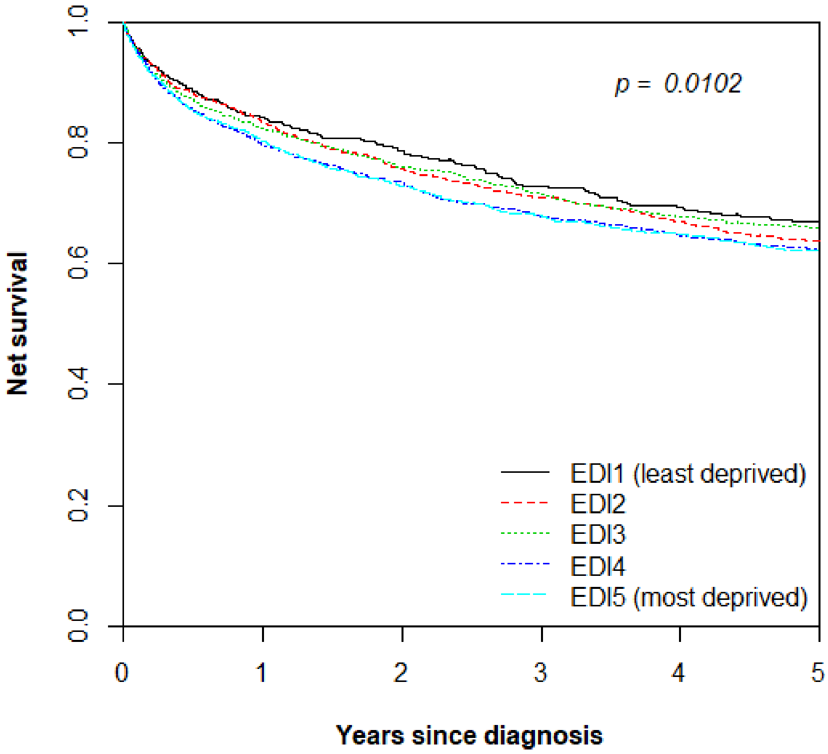

Figure 1 – Net survival by EDI category for the full cohort.

Dealing with missing information on covariates of excess hazard models - making the imputation model compatible with the substantive model

S1 - Supplementary Material

Excess hazard model with missing data on covariates

The work of Bartlett and colleagues [1] on multiple imputation using substantive model compatible fully conditional specification (SMC-FCS) was extended here to cope with excess hazard models.

We suppose that the interest lies in the time T to death from any cause. Considering T_E to be time to death from cancer and T_P the time to death from other causes, T will be the minimum between both, $T = min(T_E, T_P)$. We consider that this time T can be censored, meaning that the event of interest is not observed for all patients during the follow-up period. Let C denote the censoring time. Let $W = min(T, C)$ and $\delta = 1(T < C)$ the event indicator. We assume that T_E depends on a set of fully observed variables $Z = (Z_1, \cdots, Z_q)$ and a set of partially observed variables $X = (X_1, \cdots, X_p)$ and that T_P depends on a set of fully observed demographic variables D, considered here as a subset of Z. We assume T_E and T_P to be conditionally independent given D. Also, censoring is assumed to be noninformative.

In this relative survival setting, it is assumed that the observed hazard for one patient (λ_{O_i}) can be split in two additive hazards, the expected mortality (λ_{P_i}) and the excess hazard due to the disease in analysis (λ_{E_i}):

$$\lambda_{O_i}(t) = \lambda_{P_i}(t) + \lambda_{E_i}(t)$$

The information on expected mortality can be obtained from population life tables. The survival function for time to death from any cause is related to the observed hazard by:

$$\begin{aligned} S(t) = P(T > t) &= \exp\left[-\int_0^t \lambda_O(u)du\right] \\ &= \exp\left[-\Lambda_P(t|D) - \Lambda_E(t|X, Z)\right] \end{aligned}$$

A flexible parametric excess hazard model is assumed as the substantive model of interest:

$$\lambda_E(t|X, Z) = \lambda_0(t; \gamma) \cdot \exp[g(X_j, X_{-j}, Z; \beta)]$$

where $\lambda_E(t|X,Z)$ represents the excess hazard at time t, $\lambda_0(t;\gamma)$ represents the excess hazard baseline parametrically defined and parametrised by a set of parameters γ, and $g(X,Z;\beta)$ a function of the covariates parametrised by parameters β. The set of parameters that characterize the excess hazard function is $\psi = (\beta,\gamma)$. The method proposed by Bartlett uses a rejection sampling algorithm to draw imputations that are compatible with the substantive model. Considering first how to sample X_j for a patient for whom $\delta = 0$, assuming the time to event to be independent from time to censoring conditioned on X and Z, we have:

$$\begin{aligned} f(W=t,\delta=0|X_j,X_{-j},Z,\psi) &= f(T>t,C=t|X_j,X_{-j},Z,\psi) \\ &= P(T>t|X_j,X_{-j},Z,\psi)\cdot f(C=t|X_j,X_{-j},Z) \\ &= P(T>t|X_j,X_{-j},Z,\psi)\cdot f(C=t|Z) \\ &\le f(C=t|Z) \end{aligned}$$

The values of X_j^* are drawn from $f(X_j|X_{-j},Z,\phi_j)$ and U from an uniform distribution $U(0,1)$. X_j^* should be accepted when:

$$\begin{aligned} U &\le \frac{f(W=t,\delta=0|X_j^*,X_{-j},Z,\psi)}{f(C=t|Z)} \\ &= P(T>t|X_j^*,X_{-j},Z,\psi) \\ &= \exp\left[-\Lambda_P(t|D) - \Lambda_E(t|X_j^*,X_{-j},Z)\right] \\ &= \exp\left[-\Lambda_P(t|D) - \int_0^t \left(\lambda_0(u;\gamma)\cdot e^{g(X_j^*,X_{-j},Z;\beta)}\right) du\right] \end{aligned}$$

Assuming no time-dependent effects on covariates, this simplifies to:

$$\begin{aligned} U &\le \exp\left[-\Lambda_P(t|D) - e^{g(X_j^*,X_{-j},Z;\beta)}\cdot\Lambda_0(t;\gamma)\right] \\ &= \exp\left[-\Lambda_P(t|D)\right]\cdot\exp\left[-e^{g(X_j^*,X_{-j},Z;\beta)}\cdot\Lambda_0(t;\gamma)\right] \end{aligned}$$

For a patient who is not censored ($\delta = 1$), we have:

$$\begin{aligned} f(W=t,\delta=1|X_j,X_{-j},Z,\psi) &= f(T=t,C>t|X_j,X_{-j},Z,\psi) \\ &= f(T=t|X_j,X_{-j},Z,\psi)\cdot P(C>t|X_j,X_{-j},Z) \\ &= f(T=t|X_j,X_{-j},Z,\psi)\cdot P(C>t|Z) \\ &= \lambda(t|X_j,X_{-j},Z,\psi)\cdot P(T>t|X_j,X_{-j},Z,\psi)\cdot P(C>t|Z) \end{aligned}$$

and, since T represents time to death from any cause:

$$
\begin{aligned}
&f(W=t,\delta=1|X_j,X_{-j},Z,\psi)\\
&=[\lambda_P(t|D)+\lambda_E(t|X_j,X_{-j},Z)]\cdot P(T>t|X_j,X_{-j},Z,\psi)\cdot P(C>t|Z)\\
&=[\lambda_P(t|D)+\lambda_E(t|X_j,X_{-j},Z)]\cdot \exp\left[-\Lambda_P(t|D)-\Lambda_E(t|X_j,X_{-j},Z)\right]\cdot P(C>t|Z)\\
&=\left[\lambda_P(t|D)+\lambda_0(t;\gamma)\cdot e^{g(X_j,X_{-j},Z;\beta)}\right]\cdot\\
&\qquad \exp\left[-\Lambda_P(t|D)-\Lambda_0(t;\gamma)\cdot e^{g(X_j,X_{-j},Z;\beta)}\right]\cdot P(C>t|Z)
\end{aligned}
$$

To find the maximum the expression

$$\left[\lambda_P(t|D)+\lambda_0(t;\gamma)\cdot e^{g(X_j,X_{-j},Z;\beta)}\right]\cdot \exp\left[-\Lambda_P(t|D)-\Lambda_0(t;\gamma)\cdot e^{g(X_j,X_{-j},Z;\beta)}\right]$$

can take, we differentiated the expression with respect to g and set it to zero, resulting that the maximum of the expression is obtained when:

$$\exp(g(X_j,X_{-j},Z;\beta))=\frac{1}{\Lambda_0(t;\gamma)}-\frac{\lambda_P(t|D)}{\lambda_0(t;\gamma)}$$

Therefore,

$$
\begin{aligned}
f(W=t,\delta=1|X_j,X_{-j},Z,\psi)\leq&\frac{\lambda_0(t;\gamma)}{\Lambda_0(t;\gamma)}\cdot exp\left[-\Lambda_P(t|D)-1+\frac{\Lambda_0(t;\gamma)\cdot\lambda_P(t|D)}{\lambda_0(t;\gamma)}\right]\\
&\cdot P(C>t|Z)
\end{aligned}
$$

We can thus draw X_j^* from $f(X_j|X_{-j},Z,\phi_j)$ and $U\sim U(0,1)$, and accept X_j^* when

$$
\begin{aligned}
U\leq&\frac{f(W=t,\delta=1|X_j^*,X_{-j},Z,\psi)}{\frac{\lambda_0(t;\gamma)}{\Lambda_0(t;\gamma)}\cdot exp\left[-\Lambda_P(t|D)-1+\frac{\Lambda_0(t;\gamma)\cdot\lambda_P(t|D)}{\lambda_0(t;\gamma)}\right]\cdot P(C>t|Z)}\\
=&\frac{\left[\lambda_P(t|D)+\lambda_0(t;\gamma)\cdot e^{g(X_j^*,X_{-j},Z;\beta)}\right]\cdot exp\left[-\Lambda_P(t|D)-\Lambda_0(t;\gamma)\cdot e^{g(X_j^*,X_{-j},Z;\beta)}\right]}{\frac{\lambda_0(t;\gamma)}{\Lambda_0(t;\gamma)}\cdot exp\left[-\Lambda_P(t|D)-1+\frac{\Lambda_0(t;\gamma)\cdot\lambda_P(t|D)}{\lambda_0(t;\gamma)}\right]}
\end{aligned}
$$

References

[1] Jonathan W Bartlett, Shaun R Seaman, Ian R White, and James R Carpenter. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical methods in medical research*, 24(4), 2015.

Dealing with missing information on covariates of excess hazard models - making the imputation model compatible with the substantive model

S2 - Supplementary Material

Table S2.1 - Comparison of observed extent of disease distribution vs. distribution of Imputed values for FCS and SMC-FCS approaches.

	Localised	Local spread	Regional spread	Advanced
observed	17,2%	27,3%	31,5%	24,0%
FCS	16,6%	26,6%	29,5%	27,3%
SMC-FCS	16,3%	27,0%	29,4%	27,3%

Table S2.2 - Mean survival time (in years) by imputed extent of disease for FCS and SMC-FCS approaches.

Imputed extent	FCS	SMC-FCS
Localised	5,43	4,79
Local spread	5,07	5,08
Regional spread	4,59	4,76
Advanced	1,70	1,88

Chapter 4

Discussion and Conclusions

4.1 Discussion

The real world application that motivated the studies presented in this thesis was the evaluation of socioeconomic inequalities in survival from cancer. Several statistical methodological questions arose from this research question. Figure 4.1 summarises the integration between the several questions analysed. The evaluation of the association between

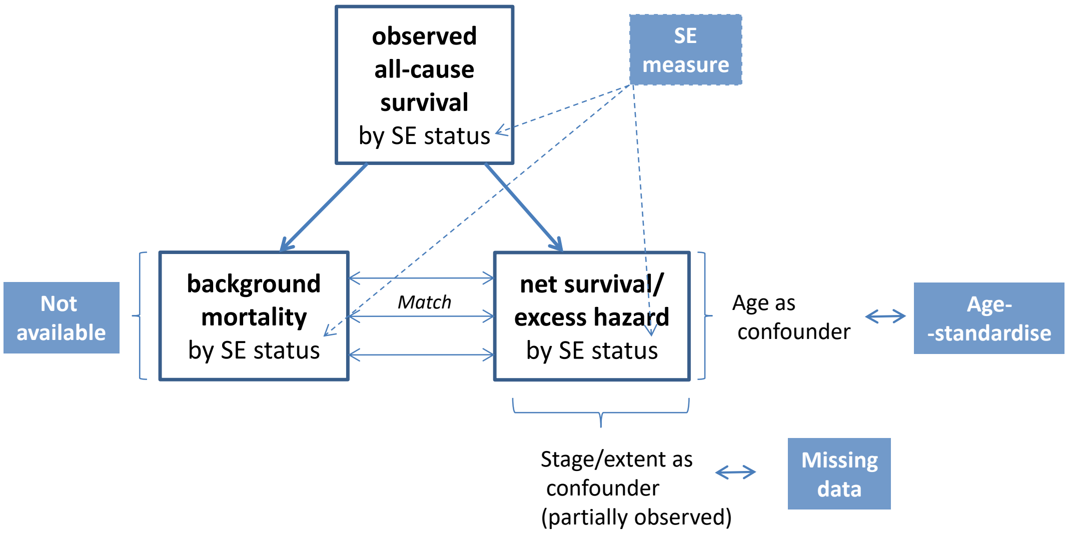
observed all-cause survival by SE status
SE measure
Not available
background mortality by SE status
Match
net survival/ excess hazard by SE status
Age as confounder
Age- -standardise
Stage/extent as confounder (partially observed)
Missing data

Figure 4.1: Schematic representation of the research questions analysed in this thesis.

patient's socioeconomic (SE) status and survival from cancer was performed in the relative survival framework. This approach was justified by the use of population-based data for which cause of death was not known or was unreliable. In this setting, the observed quantity is the all-cause survival/hazard. It is assumed that this observed hazard can be decomposed in two additive components, the hazard due to the disease under study and the hazard due to other causes. The hazard due to other causes is estimated from the population mortality, assuming that the specific disease contribution to the overall mortality is negligible. The quantity of interest, the excess hazard, can thus be estimated being the all-cause and the background hazard known. Net survival is directly related to the excess hazard. The correct estimation of the excess hazard relies on the assumption that the matched population shares the same demographic characteristics as the patients under study. Failure to meet this assumption can lead to biased excess hazard and consequently net survival estimates. In this thesis, we emphasise that when evaluating the excess hazard/net survival by SE condition, the background mortality should also be considered stratified by SE since these factors can affect both excess and other causes hazard. Since SE-specific life tables were not available for Portugal, this question had to be addressed. In the first study evaluating SE inequalities presented in Section 3.2, a sensitivity analysis to the choice of life tables was performed. Life tables stratified by SE status were then built for Portugal (Section 3.3) and applied in the estimation of net survival from cancer in Section 3.4.

Net survival from most cancers depends on age. When comparing two different populations (corresponding to different regions, periods or, for instance, SE subgroups of the same population) that have the same age-specific survival probabilities, the crude net survival in both populations can differ due to differences in the age-structure between them. To allow comparability between those populations age-standardised measures should be used. This question was addressed with special focus on situations where data are sparse in Section 3.1.

Socioeconomic inequalities in survival from cancer can partially be explained by earlier diagnosis in some SE groups relatively to others. It was thus important to adjust for stage/extent of disease at diagnosis when evaluating the association between SE factors and cancer survival. The information on this key prognostic factor had a considerable proportion of missing information in the database used in these studies. Several factors contributed to this missingness and could be related to patient's characteristics or to the registration process itself. Population-based cancer registries receive information from a large number of sources making most of the times hard to retrieve this type of information retrospectively. Statistical methods were then needed to deal with missing data that could take advantage of the available information and not only of the information from complete cases. Existing methods to deal with missing information on covariates were extended to be used with excess hazard models (Section 3.4).

No information on SE factors at individual level was available for the patients analysed. How to attribute SE status to the cancer patients and how to stratify background mortality by SE status were questions that also needed to be addressed along the several studies presented.

Summarizing, and considering the motivating application, the main research questions identified and analysed in this thesis were: evaluation of methods to estimate age-standardised net survival; analysis and extension of methods to model the excess hazard function in the presence of missing data on covariates; development of deprivation-specific life-tables to allow comparability of background mortality between the cancer cohort and the population comparison group; assessment of socioeconomic inequalities in survival from cancer. Below, some considerations regarding the several studies are addressed.

In population-based cancer survival, age standardisation is performed using a discretisation of the age distribution into age groups. The calculation of the standardised net survival results thus from a weighted average of age group-specific net survivals. These specific estimates can be obtained using a non-parametric estimator or using a model-based predictions. Either way, the age group estimates will depend on the age distribution within each age group of the cancer patients in the cohort being analysed. When age is considered as a continuous variable and the excess hazard is modelled with flexible functions, net survival of each individual can be thinly predicted. The net survival of a given age group is a weighted average of age-specific predicted survivals where the weights are given by the sample age distribution itself. When the data are sparse, this leads to unstable net survival estimates even if the model allows to smoothly predict exact individual net survivals. Furthermore, it is possible that some age groups have no observations making it impossible to estimate age group specific survival. An alternative model-based approach to estimate survival for each age group, prior to performing the classical age-standardisation, was thus proposed and evaluated. In this approach, instead of using the sample age distribution, survival was predicted from the model in a reference age, external to the sample age distribution. In this way, the estimate was no longer dependent on the sample age distribution or availability of data in each subgroup of patients. Considering the age group-specific survival given by these predictions in reference ages corresponds to making an external standardisation complementary to the classical standardisation using the Corazziari weights. In the study developed, the common number of age groups and weights were used. Other alternative ways of estimating an age-standardised measure can however be thought. The number of age groups or the age point (or points) where survival is estimated in each age group can be different. In the SUDCAN study, the net survival within each age group was calculated by averaging the net survivals predicted from the model for each annual age using the age weights within the age-class as observed over the entire data (country and site specific) [84]. This standardisation allowed comparisons between years of diagnosis since the age structure was constant over time. Nevertheless, it was a standardisation specific of that particular study directed at analysing trends in survival in each country and not at making survival comparisons between regions.

As alternative, considering that the model that allows age specific survival probabilities to be predicted is known (after having been fitted to the data), one might consider a standard population finely specified by individual age. This would use the entire information that could be extracted from the model (by considering the full survival by age profile). Also, only one weighted average would be used instead of standardising in two steps (within each group and then over age groups). Simulations are still needed to evaluate if this approach would lead to an estimator of age-standardised net survival with better performance in terms of bias, empirical coverage and mean square error than the proposed approach in Section 3.1.

Although net survival predictions can be obtained from an excess hazard model for any age, even for ages not observed in a particular sample, prior to making predictions the model has to be fitted to the data. When samples are small or the data are sparse, specially for low or high ages, the model fit can be poor. In the simulation study performed, a large variability in the models fitted was obtained. Consequently, a large variability in the survival predictions was observed. Also, using a stepwise algorithm to select the model that best fits the data, a simpler model than the 'true' one tended to be chosen. Due to small sample sizes and consequently low power of the test involved, statistically significant non-linear or time-dependent effects were hard to detect. This issue of poor model fitness in small samples is however transversal to any approach that could be used for producing model-based age-standardised net survival estimates.

The association between socioeconomic status and net survival from colorectal cancer was evaluated in studies II and IV. In the first study a cohort of patients diagnosed in the North region of Portugal in the period 2000-2002 was analysed. In the last study, the cohort analysed was from patients diagnosed a decade later (2010-2012). Since individual information on SES is not routinely available in Portuguese Cancer Registries, neither it was possible to link cancer patients to any data source with that type of individual information, area-level indicators were used. In the first of these two studies, a simple indicator (education) and a composite index (EDI) were used. Surprisingly, although education also contributed to the construction of the EDI, education alone seemed to better discriminate results in survival in the different deprivation groups. In Study II, the SES effect was only adjusted for age and sex. When this study was performed, the adjusted life tables were not yet available. Instead, general life tables, i.e. not stratified by SES, were used. Having the conscience that this option could result in overestimation of the deprivation gap, a sensitivity analysis to variations in background mortality was performed. In this analysis, the deprivation specific life tables built for England were used as basis. Although the SES disparities in overall mortality in England can differ from the ones in Southern countries as Portugal, those were used since this type of information is not available for most European countries. This sensitivity analysis showed that the inequalities in survival from cancer vanished even if the differences in mortality between SES groups in Portugal was a relatively small fraction of the differences observed in England. To allow a finer adjustment of the background mortality, deprivation-specific life tables were built for Portugal (Study III). Again, no information at the individual level on SES was available for each deceased person. To allow comparability between cancer patients and population mortality, the same index that was used to classify cancer patients was used to discriminate between different SES groups in the general population. In studies III and IV, the option fell in using only the European Deprivation Index. This has the advantage of being an index available for different European countries, allowing larger comparability between the studies developed and other studies done for different regions.

Socioeconomic inequalities in overall mortality were found for Portugal. These were larger in men than in women. When evaluating SES inequalities in cancer survival, it was thus expectable that the use of deprivation-specific life tables (instead of the general ones) would have a larger impact in men than in women.

In study IV, the evaluation of SES in survival from cancer used the deprivation-specific life tables built in the previous study. In this case, each patient was matched to a population subgroup that shared the same socioeconomic condition, enabling a more exact estimation of their net survival. Besides adjusting for age and sex, the effect of SES in survival was also adjusted for extent of disease at diagnosis. This variable is the major prognostic factor of cancer survival. It was thus important to adjust for it to understand if potential SES inequalities are explained (or at least partially explained) by earlier diagnosis in one SES groups relatively to others.

The extent of disease at diagnosis had a large proportion of missing data. In the cohort analysed this proportion was around 40%. With the developments in the cancer registration processes that are foreseen for the future, it is expectable that the proportion of missing data will tend to reduce. But for the moment, and for retrospective analysis the occurrence of missing data remains an important issue.

The most common method to handle missing data in the context of cancer survival analysis is multiple imputation. Although the application of this method using standard approaches can be found in the literature, when the substantive model of interest is an excess hazard model this can produce biased results as shown in the simulation study presented in Section 3.4.

Three approaches were used to handle with the missing information on extent of disease in the evaluation of SE inequalities. The complete-case analysis produces unbiased estimates if the missing mechanism is MCAR. This assumption is however untestable so it is not possible to know in advance if that type of approach is valid for the data in hand. The multiple imputation approach lies on the assumption that the missing mechanism is MAR. Including in the imputation models as many predictors as possible increases the plausibility of the assumption that missingness only depends on observed information [27]. In the application analysed, the number of variables available for being both predictors of values being missing and of the underlying unseen values was limited. If more variables associated with extent of disease were available, the efficiency of the imputation process could have increased.

In the standard FCS algorithm, the outcome variables must be included in the imputation model. Otherwise the association between the outcome and the explanatory variables will be biased towards the null [170]. In the SMC-FCS approach, the outcome variables are not included in the imputation models. The algorithm draws proposal imputed values from the imputation model and then the specified outcome model is used to reject or accept each proposed imputed values. While in the standard MI approach there is still some arguing on how the outcome of a survival model should be included in the imputation model, this is not an issue in the SMC-FCS approach.

The practical use of the SMC-FCS algorithm may be hampered due to its high computational time. Since each imputed value must be checked for compatibility with the substantive model the process is slow specially if the proportion of rejected values is high.

4.2 Data considerations

The major limitations identified in the studies developed in this thesis are related to data availability. The evaluation of socioeconomic inequalities in survival from colorectal cancer (Study II) implied a major effort on getting patient's addresses and its subsequent geocoding. The years of diagnosis of the patients included in this study (2000-2002) reported to a period where the completeness and quality of the information on patient's addresses registered in the Cancer Registry (CR) was inferior to what is available for more recent periods. In order to complete the information available in the CR records, an exhaustive search in the National Health Service database (RNU - *Registo Nacional de Utentes*) was performed. Some misclassification of patient's addresses could have occurred since not all patient's used to timely report address changes.

The construction of deprivation-specific life tables relied on the distribution of deaths and population by parish. The matching of the SE condition and the number of deceased individuals were based in this geographical unit. More accurate results could have been obtained with smaller geographical units. Due to the high population size of some parishes and the possible heterogeneity of SE distribution within each parish, some dilution effect could have occurred leading to some underestimation of overall mortality differences between socioeconomic groups. However, this was the smallest geographical area for which the national statistics office (INE - *Instituto Nacional de Estatística*) made the data available and it was not possible to obtain more disaggregated data.

The SES index used in both studies II and IV was the Portuguese version of the EDI. This was built based on information from the 2001 Census. An update of this index using information from the 2011 Census is being built. However, by the time Study IV (which period of diagnosis was 2010-2012) was performed, it was not available. Nevertheless, the classification of each geographical area in deprivation quintiles is not expected to undergo significant changes.

4.3 Future work

The question of age-standardisation of net survival was analysed with particular emphasis on situations with sparse data. This analysis can be extended to a broader setting with 'infinite' data. Different alternatives for standardising, as discussed above, should be tested and compared in order to evaluate which approach presents better performance.

Also, according to the methodology chosen, international standards for the age distribution and respective weights should be proposed to allow a broad use of the proposed methodology.

Concerning the methods to handle missing data, the substantive model compatible methodology was extended for excess hazard models when the covariates have linear and proportional hazards effects. This methodology should be further extended to accommodate non-linear and time-dependent effects in excess hazard models.

Socioeconomic inequalities were evaluated for colorectal cancer patients diagnosed in the North region of Portugal. The analysis can be extended to other pathologies. Although no SE disparities were found for the disease studied, the reality can be different for other cancers.

4.4 Final conclusions

Several methodological questions regarding the statistical analysis of population-based cancer survival were addressed in this thesis. Also, a real world question regarding the evaluation of socioeconomic inequalities in survival from cancer was analysed. Summarising, the main contributions of the developed work were:

- Methods to age-standardised net survival were studied and an alternative model-based approach was proposed.

- Multiple imputation methods that guarantee the compatibility between the imputation and substantive models were extended to accommodate excess hazard models.

- Deprivation-specific life tables were built for Portugal using multivariable flexible models. These life tables can be used for monitoring inequalities and in future studies that require background mortality information in the estimation of deprivation-specific net survival from any specific disease.

- The methodology to perform evaluations of socioeconomic inequalities in survival from cancer for patients was set-up. For the first time, this evaluation was performed for patients diagnosed in the North region of Portugal.

From the studies developed, the following main conclusions were draw:

- The best method to age-standardise net survival is still an open question. It has been shown that the proposed method can be a valid alternative to the conventional methods, specially in the presence of sparse data.
- The standard multiple imputation methods to handle missing data in excess hazard models with missing information on covariates can have a poor performance. The developed extension of the SMC-FCS algorithm for this context presented higher performance.
- Persistent socioeconomic inequalities in overall mortality were found for Portugal, being these larger in men than in women.
- No evidence of consistent socioeconomic inequalities in survival from colorectal cancer for patients diagnosed in the North region of Portugal were found.

The research developed along this thesis can and should be used as starting points for further research.

Bibliography

[1] Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. European Journal of Cancer. 2004 oct;40(15):2307–16.

[2] Galobardes B, Lynch J, Smith GD. Measuring socioeconomic position in health research. British Medical Bulletin. 2007;81-82(1):21–37.

[3] Woods L, Rachet B, Coleman M. Origins of socio-economic inequalities in cancer survival: a review. Annals of Oncology. 2006 jan;17(1):5–19.

[4] Aarts MJ, Lemmens VEPP, Louwman MWJ, Kunst AE, Coebergh JWW. Socioeconomic status and changing inequalities in colorectal cancer? A review of the associations with risk, treatment and outcome. European Journal of Cancer. 2010;46(15):2681–2695.

[5] Manser CN, Bauerfeind P. Impact of socioeconomic status on incidence, mortality, and survival of colorectal cancer patients: a systematic review. Gastrointestinal Endoscopy. 2014;80(1):42–60.e9.

[6] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al.. Global Cancer Observatory: Cancer Today.; 2018. Available from: `https://gco.iarc.fr/today/home`.

[7] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011 dec;45(3):1–67.

[8] Kleinbaum DG. Survival Analysis. A Self-Learning Text. New York, NY: Springer New York; 1996.

[9] Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. Springer; 2003.

[10] Collett D. Modelling survival data in medical research. Chapman & Hall/CRC; 2003.

[11] Hosmer DW, Lemeshow S, May S. Applied survival analysis: regression modeling of time-to-event data. Wiley-Interscience; 2008.

[12] Carvalho MS, Andreozzi VL, Codeço CT, Campos DP, Barbosa MTS, Shimakura SE. Análise de Sobrevivência. Teoria e aplicações em saúde. 2nd ed. Rio de Janeiro: Fiocruz; 2011.

[13] Correia F, Gouveia S, Felino AC, Costa AL, Almeida RF. Survival Rate of Dental Implants in Patients with History of Periodontal Disease: A Retrospective Cohort Study. The International Journal of Oral & Maxillofacial Implants. 2017;32(4):927–934.

[14] Marques-Alves P, Baptista R, Marinho da Silva A, Pêgo M, Castro G. Real-world, long-term survival of incident patients with pulmonary arterial hypertension. Revista Portuguesa de Pneumologia (English Edition). 2017 may;23(3):124–131.

[15] Pires-Luis AS, Vieira-Coimbra MM, Vieira FQ, Costa-Pinheiro P, Silva-Santos R, Dias PC, et al. Expression of histone methyltransferases as novel biomarkers for renal cell tumor diagnosis and prognostication. Epigenetics. 2015 oct;10(11):1033–1043.

[16] Junqueira-Neto S, Vieira FQ, Montezuma D, Costa NR, Antunes L, Baptista T, et al. Phenotypic impact of deregulated expression of class I histone deacetylases in urothelial cell carcinoma of the bladder. Molecular Carcinogenesis. 2015 jul;54(7):523–31.

[17] Correia M, Magalhães R, Silva MR, Matos I, Silva MC. Stroke Types in Rural and Urban Northern Portugal: Incidence and 7-Year Survival in a Community-Based Study. Cerebrovascular Diseases Extra. 2013;3(1):137–149.

[18] Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958 jun;53(282):457–481.

[19] Cox DR. Regression Models and Life-Tables. WileyRoyal Statistical Society; 1972.

[20] Berkson J. The calculation of survival rates. In: Wlaters W, Gray H, Priestly J, editors. Carcinoma and other malignant lesions of the stomach. Philadelphia: Sanders; 1942. p. 467–484.

[21] Pokhrel A, Hakulinen T. Age-standardisation of relative survival ratios of cancer patients in a comparison between countries, genders and time periods. European Journal of Cancer. 2009 mar;45(4):642–7.

[22] Perme MP, Stare J, Estève J. On Estimation in Relative Survival. Biometrics. 2012;68(1):113–120.

[23] 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. Statistics in Medicine. 2007;26(December 2005):2214–2228.

[24] Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. Statistics in Medicine. 2007;26(30):5486–5498.

[25] 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. Statistics in Medicine. 2016;35(2016):3066–84.

[26] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36

cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2018 sep;68(6):394–424.

[27] Nur U, Shack LG, Rachet B, Carpenter JR, Coleman MP. Modelling relative survival in the presence of incomplete data: a tutorial. International Journal of Epidemiology. 2010 feb;39(1):118–128.

[28] Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2002.

[29] Carpenter JR, Kenward MG. Missing data in randomised controlled trials a practical guide. Birmingham: Health Technology Assessment Methodology Programme; 2007.

[30] Rubin DB, Wiley InterScience. Multiple imputation for nonresponse in surveys. Wiley; 1987.

[31] Giorgi R, Belot A, Gaudart J, Launoy G, French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. Statistics in Medicine. 2008 dec;27(30):6310–6331.

[32] Falcaro M, Nur U, Rachet B, Carpenter JR. Estimating Excess Hazard Ratios and Net Survival When Covariate Data Are Missing. Epidemiology. 2015 may;26(3):421–428.

[33] Falcaro M, Carpenter JR. Correcting bias due to missing stage data in the nonparametric estimation of stage-specific net survival for colorectal cancer using multiple imputation. Cancer Epidemiology. 2017 jun;48:16–21.

[34] Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical Methods in Medical Research. 2015;24(4).

[35] Kogevinas M, Porta M. Socioeconomic differences in cancer survival: a review of the evidence. IARC scientific publications. 1997;15(138):177–206.

[36] Li R, Daniel R, Rachet B. How much do tumor stage and treatment explain socioeconomic inequalities in breast cancer survival? Applying causal mediation analysis to population-based data. European Journal of Epidemiology. 2016 jun;31(6):603–611.

[37] Woods LM, Rachet B, Coleman MP. Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. British Journal of Cancer. 2005;92(7):1279–1282.

[38] Carstairs V, Morris R. Deprivation and mortality: an alternative to social class? Community Medicine. 1989 aug;11(3):210–9.

[39] Townsend P, Phillimore P, Beattie A. Health and deprivation: inequality and the North. Croom Helm; 1988.

[40] Guillaume E, Pornet C, Dejardin O, Launay L, Lillini R, Vercelli M, et al. Development of a cross-cultural deprivation index in five European countries. Journal of Epidemiology and Community Health. 2015 dec;p. jech–2015–205729.

[41] Zadnik V, Guillaume E, Lokar K, Žagar T, Primic Žakelj M, Launoy G, et al. Slovenian Version of The European Deprivation Index at Municipal Level. Zdravstveno Varstvo. 2018 jun;57(2):47–54.

[42] Pornet C, Delpierre C, Dejardin O, Grosclaude P, Launay L, Guittet L, et al. Construction of an adaptable European transnational ecological deprivation index: the French version. Journal of Epidemiology and Community Health. 2012 nov;66(11):982–9.

[43] Ribeiro S, Furtado C, Pereira J. Associação entre as doenças cardiovasculares e o nível socioeconómico em Portugal. Revista Portuguesa de Cardiologia. 2013 nov;32(11):847–854.

[44] Oliveira CM, Alves SM, Pina MF. Marked socioeconomic inequalities in hip fracture incidence rates during the Bone and Joint Decade (2000–2010) in Portugal: age

and sex temporal trends in a population based study. Journal of Epidemiology and Community Health. 2016 aug;70(8):755–763.

[45] Santos J, Kislaya I, Antunes L, Santos AJ, Rodrigues APP, Neto M, et al. Diabetes: Socioeconomic Inequalities in the Portuguese Population in 2014. Acta Médica Portuguesa. 2017 aug;30(7-8):561–567.

[46] Ribeiro AI, Mayer A, Miranda A, Pina MF. The Portuguese Version of the European Deprivation Index: An Instrument to Study Health Inequalities. Acta Médica Portuguesa. 2017;30(1):17.

[47] Sudhakar A. History of Cancer, Ancient and Modern Treatment Methods. Journal of Cancer Science & Therapy. 2009 dec;1(2):1–4.

[48] RORENO. Registo Oncológico Nacional 2010. Instituto Português de Oncologia do Porto Francisco Gentil - EPE; 2016.

[49] Stewart BW, Wild CP. World cancer report 2014. World Health Organization. 2014;p. 1–2.

[50] Gospodarowicz M, O'Sullivan B. Prognostic factors in cancer. Seminars in Surgical Oncology. 2003 jan;21(1):13–18.

[51] Dos Santos Silva I. Interpretation of epidemiological studies. Cancer Epidemiology: Principles and Methods. 1999;p. 277–302.

[52] Portal do Instituto Nacional de Estatística;. Available from: `https://www.ine.pt/`.

[53] WHO. International Classification of Diseases for Oncology (ICD-O). 3rd ed. WHO; 2013.

[54] WHO. International statistical classification of diseases and related health problems 10th revision Volume 2 Instruction manual. 5th ed. WHO, editor; 2016.

[55] Ferlay J, Burkhard C, Whelan S, Parkin DM. Check and conversion programs for Cancer Registries. Lyon; 2005.

[56] De Angelis R, Sant M, Coleman MP, Francisci S, Baili P, Pierannunzio D, et al. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE–5 - A population-based study. The Lancet Oncology. 2014 jan;15(1):23–34.

[57] Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang XS, 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). The Lancet. 2015 mar;385(9972):977–1010.

[58] Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. The Lancet. 2018 mar;391(10125):1023–1075.

[59] Assembleia da República. Lei 53/2017. Lisboa: Diário da República n.º 135/2017, Série I de 2017-07-14; 2017.

[60] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer. 2003;89(2):232–238.

[61] Pohar Perme M, Estève J, Rachet B. Analysing population-based cancer survival settling the controversies. BMC Cancer. 2016 dec;16(1):933.

[62] Nelson W. Theory and Applications of Hazard Plotting for Censored Failure Data. Technometrics. 1972 nov;14(4):945.

[63] Aalen O. Nonparametric Inference for a Family of Counting Processes. Institute of Mathematical Statistics; 1978.

[64] Borgan Ø. Nelson-Aalen Estimator. In: Wiley StatsRef: Statistics Reference Online. Chichester, UK: John Wiley & Sons, Ltd; 2014. .

[65] Cronin KA, Feuer EJ. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. Statistics in Medicine. 2000 jul;19(13):1729–1740.

[66] Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. National Cancer Institute Monograph. 1961 sep;6:101–21.

[67] Ederer F. A simple method for determing standard errors of survival rates, with tables. Journal of Chronic Diseases. 1960 jun;11:632–45.

[68] Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. Biometrics. 1982 dec;38(4):933–42.

[69] Coleman MP, Quaresma M, Berrino F, Lutz JM, De Angelis R, Capocaccia R, et al. Cancer survival in five continents: a worldwide population-based study (CONCORD). The Lancet Oncology. 2008 aug;9(8):730–756.

[70] Berrino F, Gatta G, Chessa E, Valente F, Capocaccia R. The EUROCARE II study. European Journal of Cancer. 1998 dec;34(14):2139–2153.

[71] Hakulinen T, Seppä K, Lambert PC. Choosing the relative survival method for cancer survival estimation. European Journal of Cancer. 2011;47(14):2202–2210.

[72] Rossi S, Baili P, Capocaccia R, Caldora M, Carrani E, Minicozzi P, et al. The EUROCARE-5 study on cancer survival in Europe 1999-2007: Database, quality checks and statistical analysis methods. European Journal of Cancer. 2015 oct;51(15):2104–2119.

[73] Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: The importance of allowing for informative censoring. Statistics in Medicine. 2012;31(8):775–786.

[74] Roche L, Danieli C, Belot A, Grosclaude P, Bouvier AM, Velten M, 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. International Journal of Cancer. 2013;132(10):2359–2369.

[75] Dickman PW, Lambert PC, Coviello E, Rutherford MJ. Estimating net survival in population-based cancer studies. International Journal of Cancer. 2013 jul;133(2):519–521.

[76] Lambert PC, Dickman PW, Rutherford MJ. Comparison of different approaches to estimating age standardized net survival. BMC Medical Research Methodology. 2015 jan;15:64.

[77] Seppä K, Hakulinen T, Pokhrel A. Choosing the net survival method for cancer survival estimation. European Journal of Cancer. 2015 jun;51(9):1123–1129.

[78] Seppä K, Hakulinen T, Läärä E, Pitkäniemi J. Comparing net survival estimators of cancer patients. Statistics in Medicine. 2016 may;35(11):1866–1879.

[79] Jooste V, Grosclaude P, Remontet L, Launoy G, Baldi I, Molinié F, et al. Unbiased estimates of long-term net survival of solid cancers in France. International Journal of Cancer. 2013 may;132(10):2370–2377.

[80] Mounier M, Bossard N, Remontet L, Belot A, Minicozzi P, De Angelis R, 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. 2015 nov;2(11):e481–e491.

[81] Morris M, Woods LM, Bhaskaran K, Rachet B. Do pre-diagnosis primary care consultation patterns explain deprivation-specific differences in net survival among women with breast cancer? An examination of individually-linked data from the UK West Midlands cancer registry, national screening programme. BMC Cancer. 2017 feb;17(1):155.

[82] Cowppli-Bony A, Uhry Z, Remontet L, Voirin N, Guizard AV, Trétarre B, et al. Survival of solid cancer patients in France, 1989–2013. European Journal of Cancer Prevention. 2017 jun;26(6):461–468.

[83] Delacour-Billon S, Mathieu-Wacquant AL, Campone M, Auffret N, Amossé S, Allioux C, et al. Short-term and long-term survival of interval breast cancers taking into account prognostic features. Cancer Causes & Control. 2017 jan;28(1):69–76.

[84] Uhry Z, Bossard N, Remontet L, Iwaz J, Roche L, GRELL EUROCARE-5 Working Group and the CENSUR Working Survival Group. New insights into survival trend analyses in cancer population-based studies. European Journal of Cancer Prevention. 2017 jan;26:S9–S15.

[85] Grafféo N, Castell F, Belot A, Giorgi R. A log-rank-type test to compare net survival distributions. Biometrics. 2016 sep;72(3):760–9.

[86] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis an introduction to concepts and methods. British Journal of Cancer. 2003 aug;89(3):431–436.

[87] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. Statistics in Medicine. 1990 may;9(5):529–538.

[88] Sasieni P. Proportional excess hazards. Biometrika. 1996 mar;83(1):127–141.

[89] Perme MP, Henderson R, Stare J. An approach to estimation in relative survival regression. Biostatistics. 2009;10(1):136–146.

[90] Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, et al. A relative survival regression model using B-spline functions to model non-proportional hazards. Statistics in Medicine. 2003 sep;22(17):2767–2784.

[91] De Boor C. A practical guide to splines. Springer; 2001.

[92] Crowther MJ, Lambert PC. stgenreg: A stata package for general parametric survival analysis. Journal of Statistical Software. 2013;53(12):1–17.

[93] Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. Statistics in Medicine. 2005 dec;24(24):3871–3885.

[94] Hakulinen T, Tenkanen L. Regression Analysis of Relative Survival Rates. Apllied Statistics. 1987;36(3):309–317.

[95] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. Statistics in Medicine. 2004;23(1):51–64.

[96] Crowther MJ, Lambert PC. A general framework for parametric survival analysis. Statistics in Medicine. 2014;33(30):5280–5297.

[97] Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine. 2002 aug;21(15):2175–2197.

[98] Lambert PC, Royston P, Lambert PC, Royston P. Further development of exible parametric models for survival analysis. Stata Journal. 2009;9(2):265–290.

[99] Royston P, Lambert PC. Flexible parametric survival analysis using Stata : beyond the Cox model. Stata Press; 2011.

[100] Findley DF. Model Selection: Akaike's Information Criterion. In: Encyclopedia of Statistical Sciences. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2006. .

[101] 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. 2014;33(19):3318–3337.

[102] Stare J, Pohar M, Henderson R. Goodness of fit of relative survival models. Statistics in Medicine. 2005;24(24):3911–3925.

[103] Cortese G, Scheike TH. Dynamic regression hazards models for relative survival. Statistics in Medicine. 2008 aug;27(18):3563–84.

[104] Danieli C, Bossard N, Roche L, Belot A, Uhry Z, Charvat H, et al. Performance of two formal tests based on martingales residuals to check the proportional hazard assumption and the functional form of the prognostic factors in flexible parametric excess hazard models. Biostatistics. 2017 jul;77(3):147–160.

[105] Rabeneck L, Souchek J, El-Serag HB. Survival of colorectal cancer patients hospitalized in the Veterans Affairs Health Care System. The American Journal of Gastroenterology. 2003 may;98(5):1186–1192.

[106] Lyratzopoulos G, Sheridan GF, Michie HR, McElduff P, Hobbiss JH. Absence of socioeconomic variation in survival from colorectal cancer in patients receiving surgical treatment in one health district: cohort study. Colorectal Disease. 2004 nov;6(6):512–517.

[107] Nur U, Rachet B, Parmar MKB, Sydes MR, Cooper N, Lepage C, et al. No socioeconomic inequalities in colorectal cancer survival within a randomised clinical trial. British Journal of Cancer. 2008;99(11):1923–1928.

[108] Egeberg R, Halkjær J, Rottmann N, Hansen L, Holten I. Social inequality and incidence of and survival from cancers of the colon and rectum in a population-based study in Denmark, 19942003. European Journal of Cancer. 2008 sep;44(14):1978–1988.

[109] Hussain SK, Altieri A, Sundquist J, Hemminki K. Influence of education level on breast cancer risk and survival in Sweden between 1990 and 2004. International Journal of Cancer. 2008;122(1):165–169.

[110] Dejardin O, Remontet L, Bouvier aM, Danzon A, Trétarre B, Delafosse P, et al. Socioeconomic and geographic determinants of survival of patients with digestive cancer in France. British Journal of Cancer. 2006;95(7):944–9.

[111] Kim J, Artinyan A, Mailey B, Christopher S, Lee W, McKenzie S, et al. An interaction of race and ethnicity with socioeconomic status in rectal cancer outcomes. Annals of Surgery. 2011;253(4):647–654.

[112] Gorey KM, Luginaah IN, Bartfay E, Fung KY, Holowaty EJ, Wright FC, et al. Effects of socioeconomic status on colon cancer treatment accessibility and survival in Toronto, Ontario, and San Francisco, California, 1996-2006. American Journal of Public Health. 2011;101(1):112–119.

[113] Ueda K, Kawachi I, Tsukuma H. Cervical and corpus cancer survival disparities by socioeconomic status in a metropolitan area of Japan. Cancer Science. 2006 apr;97(4):283–291.

[114] Dalton SO, Steding-Jessen M, Gislum M, Frederiksen K, Engholm G, Schüz J. Social inequality and incidence of and survival from cancer in a population-based study in Denmark, 1994-2003: Background, aims, material and methods. European Journal of Cancer. 2008;44(14):1938–1949.

[115] Auvinen A, Karjalainen S. Possible explanations for social class differences in cancer patient survival. IARC Sci Publ. 1997;(138):377–397.

[116] Kravdal Ø. Social inequalities in cancer survival. Population Studies. 2000 jan;54(1):1–18.

[117] Gordon D. Census based deprivation indices: Their weighting and validation. Journal of Epidemiology and Community Health. 1995;49(SUPPL. 2):39–44.

[118] Eurostat. Access to Microdata. European Union Statistics on Income and Living Conditions (EU-SILC); 2015. Available from: `http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions`.

[119] Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N, et al. Trends and socioeconomic inequalities in cancer survival in England and Wales up to 2001. British Journal of Cancer. 2004 apr;90(7):1367–1373.

[120] Mitry E, Rachet B, Quinn MJ, Cooper N, Coleman MP. Survival from cancer of the rectum in England and Wales up to 2001. British Journal of Cancer. 2008 sep;99(S1):S30–S32.

[121] Pollock AM, Vickers N. Breast, lung and colorectal cancer incidence and survival in South Thames Region, 1987-1992: the effect of social deprivation. Journal of Public Health Medicine. 1997;19(3):288–94.

[122] Lejeune C, Sassi F, Ellis L, Godward S, Mak V, Day M, et al. Socio-economic disparities in access to treatment and their impact on colorectal cancer survival. International Journal of Epidemiology. 2010;39(3):710–717.

[123] Fowler H, Belot A, Njagi EN, Luque-Fernandez MA, Maringe C, Quaresma M, et al. Persistent inequalities in 90-day colon cancer mortality: an English cohort study. British Journal of Cancer. 2017 oct;117(9):1396–1404.

[124] Abdel-Rahman ME, Butler J, Sydes MR, Parmar MKB, Gordon E, Harper P, et al. No socioeconomic inequalities in ovarian cancer survival within two randomised clinical trials. British Journal of Cancer. 2014 jul;111(3):589–597.

[125] Shafique K, Morrison DS. Socio-Economic Inequalities in Survival of Patients with Prostate Cancer: Role of Age and Gleason Grade at Diagnosis. PLoS ONE. 2013 feb;8(2):e56184.

[126] Di Salvo F, Caranci N, Spadea T, Zengarini N, Minicozzi P, Amash H, et al. Socioeconomic deprivation worsens the outcomes of Italian women with hormone receptor-positive breast cancer and decreases the possibility of receiving standard care. Oncotarget. 2017 sep;8(40):68402–68414.

[127] Belot A, Remontet L, Rachet B, Dejardin O, Charvat H, Bara S, et al. Describing the association between socioeconomic inequalities and cancer survival: methodological guidelines and illustration with population-based data. Clinical Epidemiology. 2018;10:561–573.

[128] Sloggett A, Joshi H. Deprivation indicators as predictors of life events 1981-1992 based on the UK ONS longitudinal study. Journal of Epidemiology and Community Health. 1998;52(4):228–233.

[129] Diez Roux AV. The Study of Group-Level Factors in Epidemiology: Rethinking Variables, Study Designs, and Analytical Approaches. Epidemiologic Reviews. 2004 jul;26(1):104–111.

[130] NHS. Supporting Information: Lower Layer Super Output Area;.

[131] INSEE. Définition - IRIS — Insee;. Available from: https://www.insee.fr/en/metadonnees/definition/c1523.

[132] INE. Instituto Nacional de Estatistica, Censos 2011;. Available from: https://censos.ine.pt/xportal/xmain?xpid=CENSOS{\&}xpgid=censos{_}base{_}cartogr.

[133] Harris AR, Bowley DM, Stannard A, Kurrimboccus S, Geh JI, Karandikar S. Socioeconomic deprivation adversely affects survival of patients with rectal cancer. British Journal of Surgery. 2009;96(7):763–768.

[134] Mackillop WJ, Zhang-Salomons J, Groome PA, Paszat L, Holowaty E. Socioeconomic status and cancer survival in Ontario. Journal of Clinical Oncology. 1997 apr;15(4):1680–1689.

[135] Gorey KM, Holowaty EJ, Fehringer G, Laukkanen E, Moskowitz A, Webster DJ, et al. An international comparison of cancer survival: Toronto, Ontario, and Detroit, Michigan, metropolitan areas. American Journal of Public Health. 1997;87(7):1156–1163.

[136] Cavalli-Björkman N, Lambe M, Eaker S, Sandin F, Glimelius B. Differences according to educational level in the management and survival of colorectal cancer in Sweden. European Journal of Cancer. 2011;47(9):1398–1406.

[137] WHO. GHO — By category — Life tables. World Health Organization;. Available from: http://apps.who.int/gho/data/node.main.687?lang=en.

[138] HMD. Human Mortality Database;. Available from: http://www.mortality.org/.

[139] Baili P, Micheli A, Montanari A, Capocaccia R. Comparison of four methods for estimating complete life tables from abridged life tables using mortality data supplied to Eurocare-3. Mathematical Population Studies. 2005;12(4):183–198.

[140] Rachet B, Maringe C, Woods LM, Ellis L, Spika D, Allemani C. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. BMC Public Health. 2015 dec;15(1):1240.

[141] Spika D, Bannon F, Bonaventure A, Woods LM, Harewood R, Carreira H, et al. Life tables for global surveillance of cancer survival (the CONCORD programme): data sources and methods. BMC Cancer. 2017 dec;17(1):159.

[142] Ali A, Dawson SJ, Blows F, Provenzano E, Ellis I, Baglietto L, et al. Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. British Journal of Cancer. 2011 feb;104(4):693–699.

[143] Soegaard M, Olsen M. Quality of cancer registry data: completeness of TNM staging and potential implications. Clinical Epidemiology. 2012 aug;4(Supplement 2 Cancer staging):1.

[144] Luo Q, Yu XQ, Cooke-Yarborough C, Smith DP, O'Connell DL. Characteristics of cases with unknown stage prostate cancer in a population-based cancer registry. Cancer Epidemiology. 2013;37(6):813–819.

[145] Di Girolamo C, Walters S, Benitez Majano S, Rachet B, Coleman MP, Njagi EN, et al. Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013. BMC Cancer. 2018 may;18(1):492.

[146] Gurney J, Sarfati D, Stanley J, Dennett E, Johnson C, Koea J, et al. Unstaged cancer in a population-based registry: Prevalence, predictors and patient prognosis. Cancer Epidemiology. 2013 aug;37(4):498–504.

[147] Worthington JL, Koroukian SM, Cooper GS. Examining the characteristics of unstaged colon and rectal cancer cases. Cancer Detection and Prevention. 2008;32(3):251–258.

[148] Ostenfeld EB, Frøslev T, Friis S, Gandrup, Madsen, Soegaard M. Completeness of colon and rectal cancer staging in the Danish Cancer Registry, 2004–2009. Clinical Epidemiology. 2012;4(Suppl 2):17–23.

[149] Merrill RM, Sloan A, Anderson AE, Ryker K. Unstaged cancer in the United States: a population-based study. BMC Cancer. 2011 dec;11(1):402.

[150] Rubin DB. Inference and missing data. Biometrika. 1976 dec;63(3):581–592.

[151] Carpenter JR, Kenward MG. Multiple Imputation and its Application. 1st ed. Chichester, UK: John Wiley & Sons, Ltd; 2013.

[152] Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples LA. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Statistics in Medicine. 2003 feb;22(4):545–557.

[153] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010 dec;29(28):2920–2931.

[154] Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research. 2013 jun;22(3):278–295.

[155] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology. 1995 dec;142(12):1255–64.

[156] Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. British Journal of Cancer. 2004 jul;91(1):4–8.

[157] Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Canadian Medical Association Journal. 2012 aug;184(11):1265–9.

[158] Pérez A, Dennis RJ, Gil JFA, Rondón MA, López A. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Statistics in Medicine. 2002;21(24):3885–3896.

[159] Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 2006 oct;59(10):1087–1091.

[160] Rubin DB. Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse. In: Proceedings of the Survey Research Methods Section of the American Statistical Association; 1978. p. 20–34.

[161] Schafer JL. Multiple imputation: A primer. Statistical Methods in Medical Research. 1999;8(1):3–15.

[162] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine. 1999 mar;18(6):681–94.

[163] Akande O, Li F, Reiter J. An Empirical Comparison of Multiple Imputation Methods for Categorical Data. The American Statistician. 2017 apr;71(2):162–170.

[164] van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation. 2006;76(12):1049–1064.

[165] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16(3):219–242.

[166] Lee KJ, Carlin JB. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology. 2010;171(5):624–632.

[167] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ (Clinical research ed). 2009;338.

[168] Murray JS. Multiple Imputation: A Review of Practical and Theoretical Findings. Statistical Science. 2018 may;33(2):142–159.

[169] Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006 oct;59(10):1092–1101.

[170] Bartlett JW, Frost C, Carpenter JR. Multiple imputation models should incorporate the outcome in the model of interest. Brain. 2011 nov;134(11):e189–e189.

[171] Bartlett JW, Morris TP. Multiple imputation of covariates by substantive-model compatible fully conditional specification. Stata Journal. 2015;15(2):437–456(20).

[172] Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2008 jun;57(3):273–291.

[173] Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. Emerging Themes in Epidemiology. 2017;14:8.

[174] Barzi F, Woodward M. Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies. American Journal of Epidemiology. 2004 jul;160(1):34–45.

[175] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data. Journal of Clinical Epidemiology. 2003;56(1):28–37.

[176] Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Medical Research Methodology. 2010 dec;10(1):7.

[177] White IR, Royston P. Imputing missing covariate values for the Cox model. Statistics in Medicine. 2009 jul;28(15):1982–98.

[178] Ramos M, Montaño J, Esteva M, Barceló A, Franch P. Colorectal cancer survival by stage of cases diagnosed in Mallorca, Spain, between 2006 and 2011 and factors associated with survival. Cancer Epidemiology. 2016 apr;41:63–70.

[179] Woods LM, Rachet B, O'Connell D, Lawrence G, Tracey E, Willmore A, et al. Large differences in patterns of breast cancer survival between Australia and England: A comparative study using cancer registry data. International Journal of Cancer. 2009 may;124(10):2391–2399.

[180] Dejardin O, Rachet B, Morris E, Bouvier V, Jooste V, Haynes R, et al. Management of colorectal cancer explains differences in 1-year relative survival between France

and England for patients diagnosed 1997-2004. British Journal of Cancer. 2013 mar;108(4):775–83.

[181] Dejardin O, Jones AP, Rachet B, Morris E, Bouvier V, Jooste V, et al. The influence of geographical access to health care and material deprivation on colorectal cancer survival: Evidence from France and England. Health and Place. 2014;30:36–44.

[182] Walters S, Maringe C, Butler J, Rachet B, Barrett-Lee P, Bergh J, et al. Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: A population-based study. British Journal of Cancer. 2013;108(5):1195–1208.

[183] Le Guyader-Peyrou S, Orazio S, Dejardin O, Maynadié M, Troussard X, Monnereau A. Factors related to the relative survival of patients with diffuse large B-cell lymphoma in a population-based study in France: does socio-economic status have a role? Haematologica. 2017;102(3):584–592.

[184] Monteiro LS, Antunes L, Santos LL, Bento MJ, Warnakulasuriya S. Survival probabilities and trends for lip, oral cavity and oropharynx cancers in Northern Portugal in the period 2000-2009. Ecancermedicalscience. 2018;12:855.

Appendices

Appendix A

Oral and Posters communications

In this appendix, the abstracts of the oral and posters communications that have been done during this thesis development are presented. The title of the communications and respective conference are presented in chronological order, starting with the oral communications and followed by the poster presentations.

A.1 Oral communications

Imputação múltipla - Uma aplicação ao tratamento de dados omissos em análise de sobrevivẽncia de doentes oncológicos
Luís Antunes, Maria José Bento, Denisa Mendonça
XIX Congresso Anual da Sociedade Portuguesa de Estatstica - Nazaré, Portugal - 2011

Socio-economic inequalities in stomach cancer survival: the importance of accounting properly for missing data
Luís Antunes, Bernard Rachet, Maria José Bento, Denisa Mendonça
XXXVII GRELL Meeting - Porto, Portugal - 2012

Assessment of socioeconomic inequalities in stomach cancer survival in the North Region of Portugal
Luís Antunes, Bernard Rachet, Maria de Fátima Pina, Maria José Bento, Denisa Mendonça
European Congress of Epidemiology - Porto, Portugal - 2012
Abstract published in: European Journal of Epidemiology (2012) 27:S1S197

Socioeconomic inequalities in bladder cancer survival in the North Region of Portugal, 1999-2006
Luís Antunes, Maria José Bento, Clara Castro, Bernard Rachet, Denisa Mendonça
XXXVIII GRELL Meeting - Siracusa, Italy - 2013

Estimation of age-standardized net survival in sparse data using a modelling approach
Luís Antunes, Denisa Mendonça, Aurélien Belot, Bernard Rachet
One-day Workshop on Survival Analysis, Lisbon, Portugal - 2017

Deprivation-specific life tables using multivariable flexible modelling - trends from 2000-2002 to 2010-2012
Luís Antunes, Denisa Mendonça, Ana Isabel Ribeiro, Camille Maringe, Bernard Rachet
III Encontro Luso-Galaico de Biometria, Aveiro, Portugal - 2018

A.2 Poster communications

Desigualdades sócio-económicas na sobrevivẽncia de doentes oncol'ogicos na presença de informação incompleta
Luís Antunes, Bernard Rachet, Maria de Fátima Pina, Maria José Bento, Denisa Mendonça
XX Congresso Anual da Sociedade Portuguesa de Estatstica - Porto, Portugal - 2012

Desigualdades sócio-económicas na sobrevivẽncia de doentes diagnosticados com tumores do estômago e bexiga na Região Norte de Portugal
Luís Antunes, Bernard Rachet, Maria José Bento, Denisa Mendonça
I Encontro Portugus de Biometria & I Encontro Luso-Galaico de Biometria - Braga, Portugal - 2013

Análise de sobrevivência com informação incompleta nas covariáveis - Estudo de simulação

Luís Antunes, Bernard Rachet, Maria José Bento, Denisa Mendonça

I Encontro Portugus de Biometria & I Encontro Luso-Galaico de Biometria - Braga, Portugal - 2013

Socioeconomic inequalities in cancer survival for the most common cancer sites in the North Region of Portugal, 20002006

Luís Antunes, Bernard Rachet, Maria José Bento, Denisa Mendonça

European Congress of Epidemiology - Aarhus, Denmark - 2013

Abstract published in: European Journal of Epidemiology (2013) 28:S1S270

Colorectal cancer survival by education level in an urban area of the North region of Portugal, 2000-2002

Luís Antunes, Maria José Bento, Clara Castro, Bernard Rachet, Denisa Mendonça

XXXIX GRELL Meeting - Genève, Switzerland - 2014

Desigualdades sócio-económicas na sobrevivẽncia de doentes com cancro colorectal na cidade do Porto

Luís Antunes, Maria Fátima Pina, Maria José Bento, Denisa Mendonça

IX CONGRESSO DA APE - Alicante, Spain - 2014

Impact of the choice of life tables on the assessment of socioeconomic inequalities in survival from colorectal cancer

Luís Antunes, Bernard Rachet, Maria José Bento, Denisa Mendonça

European Congress of Epidemiology - Maastricht, The Netherlands - 2015

Abstract published in: European Journal of Epidemiology (2015) 30:7091001

Estimation of age-standardized net survival with sparse data: taking advantage of regression models

Luís Antunes, Denisa Mendonça, Aurélien Belot, Bernard Rachet

XXIII Congresso Anual da Sociedade Portuguesa de Estatstica - Lisboa, Portugal - 2017

Imputação múltipla - Uma aplicação ao tratamento de dados omissos em análise de sobrevivência de doentes oncológicos

Luís Antunes[1], **Maria José Bento**[1] **e Denisa Mendonça**[2]

[1] RORENO - Registo Oncológico Regional do Norte, {*luis.antunes, mjbento*}*@ipoporto.min-saude.pt*
[2] ICBAS/ISPUP - Universidade do Porto, *dvmendon@icbas.up.pt*

Resumo: A existência de informação incompleta é um problema comum em muitos estudos na área da saúde. A forma mais comum de lidar com a ocorrência de dados omissos consiste em não considerar na análise os registos com informação incompleta. Esta restrição na análise pode levar a inferências com diferenças substanciais daquelas que seriam obtidas se não houvesse dados omissos. A imputação múltipla tem sido uma das formas de lidar com dados omissos no pressuposto que os dados em falta dependam apenas de informação observada. Neste trabalho apresenta-se uma aplicação da imputação múltipla a um problema de análise de sobrevivência de doentes com cancro do pulmão.

Palavras–chave: Imputação múltipla, análise de sobrevivência, cancro do pulmão

Introdução

A existência de variáveis com informação incompleta é um problema recorrente em registos oncológicos de base populacional. A extensão da doença à data de diagnóstico, factor de prognóstico de maior importância, é uma variável para a qual a percentagem de casos sem informação tende a ser elevada. Numa análise de sobrevivência, a consideração apenas dos casos para os quais existe informação completa, pode introduzir enviesamentos nas conclusões que se retiram dessa mesma análise, especialmente se o mecanismo de omissão não for completamente aleatório. No pressuposto de que a falta de informação depende apenas de informação observada, a imputação múltipla é uma das formas propostas para lidar com este problema em estudos de sobrevivência com informação incompleta nas covariáveis [2].

Métodos

Aplicou-se a imputação múltipla por equações em cadeia [3] para gerar as observações das variáveis em falta, iterativamente, a partir da distribuição de cada uma dessas variáveis condicionada aos dados observados para outras variáveis. Vários conjuntos de dados completados foram gerados. Para cada um destes conjuntos foi ajustado um modelo de sobrevivência relativa. Neste modelo, com estrutura de modelo linear generalizado com erro de Poisson, considera-se que o risco de morte de cada paciente resulta da soma de duas componentes: uma relacionada com o risco esperado (estimado a partir de tábuas de mortalidade para a população em geral) e uma componente de excesso de risco relacionado com a doença [1]. O resultado do modelo são estimativas para razões de excesso de risco para cada covariável, ajustadas para as restantes. Os resultados obtidos para cada conjunto completado são combinados para produzir as estimativas finais. Na variância final das estimativas dos coeficientes do modelo, é tida em conta a incerteza associada aos valores estimados no processo de imputação [2].

Aplicação

Pretendeu-se estudar os factores de prognóstico mais importantes na sobrevivência de doentes de cancro do pulmão. Consideraram-se os pacientes diagnosticados no período 2000 a 2006, com idade igual ou superior a 15 anos, residentes na região Norte de Portugal à data de diagnóstico e registados no RORENO (Registo Oncológico Regional do Norte). O estadio da doença à data do diagnóstico não era conhecido em cerca de metade dos casos e a morfologia do tumor encontrava-se mal especificada em cerca de 26% dos casos. Neste trabalho, apresentam-se os resultados obtidos na modelação da sobrevivência, tendo sido usada a imputação múltipla para completar a informação nas covariáveis com informação em falta. Variáveis como estado vital, tempo de sobrevivência, idade, sexo, fonte de informação, ano de diagnóstico, base de diagnóstico, entre outras, foram usadas nos modelos de imputação. Foi efectuada uma análise comparativa entre os resultados obtidos e aqueles que se obtiveram usando apenas os casos completos.

Bibliografia

[1] Dickman, P.W., Sloggett, A., Hills, M. e Hakulinen, T (2004). Regression models for relative survival. *Statistics in Medicine*, 23, 51-64.

[2] Nur, U., Shack, L.G., Rachet, B., Carpenter, J.R. e Coleman, M.P. (2010). Modelling relative survival in the presence of incomplete data: a tutorial. *Int. J. Epidemiol*, 39(1), 118-28.

[3] Van Buuren, S., Boshuizen, H.C. e Knook, D.L. (1999). Multiple Imputation of missing blood pressure covariates in survival analysis. *Statist. Med.*, 18, 681-694.

O_26

SOCIO-ECONOMIC INEQUALITIES IN STOMACH CANCER SURVIVAL: THE IMPORTANCE OF ACCOUNTING PROPERLY FOR MISSING DATA.

Luís Antunes[1]; Bernard Rachet[2]; Maria José Bento[1]; Denisa Mendonça[3,4]

[1]Registo Oncológico Regional do Norte (RORENO), Instituto Português de Oncologia do Porto - luis.antunes@ipoporto.min-saude.pt
[2]Cancer Survival Group, London School of Hygiene and Tropical Medicine (LSHTM)
[3]Instituto Ciências Biomédicas Abel Salazar (ICBAS)
[4]Instituto Saúde Pública da Universidade do Porto (ISPUP)

INTRODUCTION

Socio-economic inequalities in cancer survival have been reported in different countries. Tumour stage may play an important role in these inequalities, but is often missing in high proportions.

OBJECTIVES

To estimate net survival from stomach cancer by deprivation, adjusted for stage and accounting for missing stage information.

MATERIAL AND METHODS

Various ecological socio-economic measures were allocated to stomach cancer patients registered by the Portuguese Institute of Oncology (Porto) in 2005-06. Up-to-five-year net survival was estimated using a flexible modelling approach enabling to model the effects of sex, age, socio-economic condition and stage. Missing data were handled using multiple imputation. We compared complete-case and imputation-based findings.

RESULTS

The analysis included 593 patients (60% male). Tumour stage was missing for about 20% of the patients, but this proportion was higher in elderly and in palliative care group. Advanced disease was more frequent in male patients, those aged 55-64 and those coming from more deprived areas. Survival of patients with known stage was significantly higher than survival in patients with unknown stage (gap: 12%).

Preliminary results show that education level was associated with inequalities in survival, although the trend was not completely clear. The difference in 5-year relative survival between the groups coming from areas with lower and higher level of education was 10%. Adjusting for tumour stage attenuated these differences. Following multiple imputation, sensitivity analysis will be performed to test the MAR assumption.

DISCUSSION AND CONCLUSIONS

This study represents one of the first attempts to study socio-economic inequalities in cancer survival in Portugal. Further studies using socio-economic indices more complete than simple indicators will be carried out. The role of stage tumour and treatment on socio-economic inequalities will be further investigated. Multiple imputations allows the use of all available information, including variables not directly considered in the analysis model, resulting in less biased and more precise estimates.

working in average 136 weeks until censoring. Socioeconomic factors were found to be associated with retirement but not with sickness absence and return to work. Contrary, previous episodes of unemployment and sickness absence were associated with the risk for sickness absence and resuming of work. Stage of disease, type of operation, ASA score and post-operative complication were all associated with the outcomes under study.

Conclusions Stage of disease, general health state of the individual (ASA score), post-operative complications and the history of sickness absence and unemployment had an impact on the transition between work, sickness absence and pension among survivors of colorectal cancer. This leads to an increased focus on the more vulnerable persons who have a history of work related problems.

OC 3.3.2

Malignant melanoma in the Arkhangelsk region, Russia in 2000–2010: epidemiology and survival

Anna Subbotina, Mikhail Valkov, Mikhail Levit, Andrej Grjibovski

International School of Public Health, Northern State Medical University, Arkhangelsk, Russia; Department of radiology and radiation oncology, Northern State Medical University, Arkhangelsk, Russia; Norwegian Institute of Public Health, Oslo, Norway

Background The incidence of malignant melanoma is increasing worldwide. The increase of incidence rates of melanoma in higher latitudes is a complex phenomenon associated with changes in both physical and lifestyle factors.

Objectives To describe incidence, mortality in Arkhangelsk region, Northwest Russia as well as to estimate survival and associated factors using the data from the Arkhangelsk Regional Cancer Registry (ARCR).

Methods Data were extracted from the ARCR. Information on population size was obtained from the Regional Bureau of Statistics. In 97 % of cases diagnosis was histologically confirmed. Mortality and incidence were estimated using all new cases registered in 2000–2010. Age-standardized mortality and incidence rates were calculated using Standard World Population. Stratified Cox Regression analysis was used for estimating survival. The potential predictors were age, sex, setting, site and stage by TNM system.

Results Altogether, 716 new cases of melanoma occurred in 2000–2010. Age of diagnosis ranged from 18 to 87 (mean 56.0) years. Women constituted 66 % and men 34 % of cases. The most common site was trunk (56 %) and legs (19 %) for men and legs (36 %) and trunk (32 %) for women. The stage distribution was: T1 15 %, T2 29 %, T3 26 %, T4 30 %; N1 10 %, N2 8 %, N3 5 % and M1 9 %. Crude incidence rate per 100,000 increased from 4.23 to 4.62 for men, and from 4.67 to 9.40 for women, and standardized incidence rate for both sexes increased from 3.82 to 5.63 from 2000 to 2010. Mortality was probably underestimated as only part of cases before 2010 were included in the registry. Standardized mortality rate increased from 0.57 to 2.23 per 100 000 in 2000–2010. The stratified Cox Regression model included age, setting, and TNM stage (stratified by sex). Significant predictors were urban setting compared to rural (HR = 0.65, 95 % CI = 0.44–0.96), stage T4 compared to T0 (HR = 2.50, 95 % CI = 1.29–4.81), stages N1 (HR = 4.00, 95 % CI = 2.47–6.48), N2 (HR = 2.95, 95 % CI = 1.64–5.30), N3 (HR = 4.31, 95 % CI = 2.03–9.18) compared to N0, and stage M1 (HR = 4.86, 95 % CI = 2.68–8.80) compared to M0.

Conclusions Incidence of malignant melanoma has increased in 2000–2010, particularly among women. Setting and stage by TNM system were significantly associated with survival when adjusted for other variables. Gender and setting differences in survival can possibly been explained by differences in lifestyle warranting further investigation.

OC 3.3.3

Assessment of socioeconomic inequalities in stomach cancer survival in the North Region of Portugal

Luis Antunes, Bernard Rachet, Maria de Fátima Pina, Maria José Bento, Denisa Mendonça

Registo Oncológico Regional do Norte, Instituto Português de Oncologia, Porto (RORENO); Cancer Survival Group, London School of Hygiene and Tropical Medicine (LSHTM); Faculdade de Medicina da Universidade do Porto (FMUP); Instituto Ciências Biomédicas Abel Salazar (ICBAS); Instituto Saúde Pública da Universidade do Porto (ISPUP); Instituto Nacional de Engenharia Biomédica (INEB)

Background Cancer survival is known to be associated with socioeconomic factors. Several studies performed in different countries have demonstrated socioeconomic inequalities. They are more evident for cancers that have a better prognosis and for which treatment and possibility of cure exists. Several factors can contribute for explaining those differences in survival. However, information on some of these factors, such as stage or morphology, is commonly, incomplete. Regional cancer registries collect information from many different hospitals and pathology laboratories, which make information recovery difficult.

The aim of this study is to assess socioeconomic inequalities in stomach cancer survival in the North Region of Portugal, adjusted for stage, sex and age and accounting for missing stage information.

Materials and methods All stomach cancer patients registered in the Portuguese Institute of Oncology, diagnosed in the period 2005–2006, aged 15 years or older, were considered for analysis. Various ecological socio-economic measures were allocated to the patients, by matching patient's addresses with information from the National Statistics Office. Up-to-five-year net survival was estimated using a flexible modelling approach enabling to model the effects of sex, age, socio-economic condition and stage. Missing data were handled using multiple imputation procedures.

Results The analysis included 591 patients (60 % male). Tumour stage was missing for less than 20 % of the patients, but this proportion was higher in elderly and in palliative care group. Preliminary results showed that patients coming from areas with the lowest proportion of persons with compulsory education level had a lower survival than the remaining patients. Adjusting for tumour stage attenuated these differences.

Discussion This study represents one of the first attempts to study socio-economic inequalities in cancer survival in Portugal. We used simple socioeconomic indicators, such as education. Further studies using more complete deprivation indices should be considered in the future. The role of stage tumour and treatment on socioeconomic inequalities will be further investigated. Multiple imputation allows the use of all available information, including variables not directly considered in the survival analysis model, resulting in less biased and more precise estimates.

OC 3.3.4

Survival analysis of second primary cancers in North Portugal: a population-based registry evaluation

Luis Figueiredo, Luis Antunes, Maria José Bento, Nuno Lunet

SOCIOECONOMIC INEQUALITIES IN BLADDER CANCER SURVIVAL IN THE NORTH REGION OF PORTUGAL, 1999-2006.

Luís Antunes[1]; Maria José Bento[1]; Clara Castro[1]; Bernard Rachet[2]; Denisa Mendonça[3,4]

[1] Registo Oncológico Regional do Norte (RORENO), Instituto Português de Oncologia do Porto
[2] Cancer Survival Group, London School of Hygiene and Tropical Medicine (LSHTM)
[3] Instituto Ciências Biomédicas Abel Salazar (ICBAS)
[4] Instituto Saúde Pública da Universidade do Porto (ISPUP)

e-mail: luis.antunes@ipoporto.min-saude.pt

Introduction

Socioeconomic conditions are known to affect cancer survival although for bladder cancer, results obtained in different studies are not consensual.

Objectives

To describe the survival from bladder cancer patients diagnosed in the period 1999 to 2006 in the North Region of Portugal and to study the influence of socioeconomic conditions in survival.

Material and Methods

All malignant invasive bladder cancer patients, registered by the North Region of Portugal Cancer Registry (RORENO), with residence at diagnosis in its area of influence, diagnosed in the period 1999 to 2006, aged 15 years or older, were considered for analysis. Socioeconomic categories were assigned to each patient using ecological variables. Relative survival was calculated using Ederer II method and excess hazards were estimated using parametric flexible models.

Results

A total of 4143 patients (78% male) were diagnosed in the period of interest. After excluding cases with no follow-up information (4.3%) and with incomplete residence (2.9%), 3845 cases were included for analysis. Overall 5-year relative survival was 72.8%. Period of diagnosis was divided in two groups (1999-2002 and 2003-2006). An increase in survival was observed from the first to the more recent period (70.3% to 74.5%). Five-year relative survival ranged from 89.5% for the younger age group (15-44) to 65.0% for the oldest (75+). Women presented a worse survival (71.5%) compared to men (73.2%), although after adjusting for age, women presented a lower hazard (HR=0.84; CI95: 0.72-0.99). Patients from areas with lower level of education, higher illiteracy, higher indices of rurality and lower levels of accessibility to goods and services presented a lower survival. After adjusting for age, sex and period of diagnosis, patients from areas with the highest quintile of rurality had a hazard ratio of 1.34 (CI95: 1.08-1.66) when compared to the lowest quintile.

Discussion and conclusions

An increase in survival has been observed in the last decade for bladder cancer patients in the North Region of Portugal. Some inequalities in cancer survival were observed, specially related to the rurality of patients' area of residence. The reasons for these differences need to be further investigated.

Poster

Estimation of age-standardized net survival in sparse data using a modelling approach

Luis Antunes, Faculdade de Ciências da Universidade do Porto, antunes.lj@gmail.com

Denisa Mendonça, Instituto de Ciências Biomédicas Abel Salazar, dvmendon@icbas.up.pt

Aurelien Belot, Cancer Survival Group, London School of Hygiene and Tropical Medicine, United Kingdom, Aurelien.Belot@lshtm.ac.uk

Bernard Rachet, Cancer Survival Group, London School of Hygiene and Tropical Medicine, United Kingdom, Bernard.Rachet@lshtm.ac.uk

Keywords: Net survival, Excess hazard modelling, Cancer, Sparse data.

Summary:

Cancer survival analysis is of major importance in the evaluation of cancer care practices provided to populations. International comparison of survival probabilities from cancer should take into account differences in patient’s population age structure since survival from cancer is often age dependent. This is usually achieved through direct age-standardization using a common age distribution standard such as the International Cancer Survival Standards. The direct age-standardization implies the estimation of survival for each age group. Often, the extreme age groups (youngest or oldest, depending on the cancer) are sparse and their net survival estimates are either very unstable or even impossible to obtain a few years after diagnosis.

Net survival, the survival that would be observed in the absence of causes of death not related to the disease in study, can be estimated using the Pohar-Perme estimator or a modelling approach. If the model is correctly specified, both methods should produce the same estimate. When age is considered as a continuous variable and the excess hazard is modelled with flexible functions (e.g. splines), net survival of each individual can be thinly predicted for any time since diagnosis. The net survival of a given age group is obtained as the mean of the individual net survival of the subjects in this age group. Although a flexible modelling approach is used, net survival estimate of each age group depends on the observed number of subjects in each group as well as on their observed age-distribution. This will again lead to unstable net survival estimates when the data are sparse even if the model allows to smoothly predict exact individual net survivals. Age group-specific estimates given by the non-parametric Pohar-Perme estimator are also very unstable on such datasets.

An alternative approach to the estimation of age-standardized net survival would be to predict survival (model-based) for a reference age in each age group or for a reference age instead of averaging the individual’s survival.

The main aim of this study was to evaluate and compare methods for the estimation of age-standardized net survival when data are sparse. We compared three different approaches. Two model-based estimators of survival and the non-parametric estimator proposed by Pohar-Perme. In the first model-based approach, net survival was estimated averaging individual survivals within each age group. In the second, survival was estimated at a reference age in each age group. A flexible parametric model on the log hazard scale was used to model the excess hazard. We compared empirically the three approaches on small randomly selected samples from a large simulated dataset under different scenarios of age and year of diagnosis dependence.

DEPRIVATION-SPECIFIC LIFE TABLES USING MULTIVARIABLE FLEXIBLE MODELLING - TRENDS FROM 2000-2002 TO 2010-2012

Luís Antunes[1,2,3], Denisa Mendonça[3,4], Ana Isabel Ribeiro[3], Camille Maringe[5], Bernard Rachet[5]

[1]Department of Epidemiology, Portuguese Oncology Institute – Porto, Portugal
[2]Faculty of Sciences, University of Porto, Portugal
[3]EPIUnit – Institute of Public Health – University of Porto (ISPUP), Porto, Portugal
[4]Institute of Biomedical Sciences Abel Salazar, University of Porto, Portugal
[5]Cancer Survival Group, London School of Hygiene and Tropical Medicine, United Kingdom

ABSTRACT

Mortality data are an important indicator of population health and development. Information on socioeconomic inequalities in mortality is crucial for policy decisions. The aim of this study was to build deprivation-specific life tables using the Portuguese version of the European Deprivation Index (EDI) as a measure of area socioeconomic deprivation, and to evaluate its trends between the periods 2000-2002 and 2010-2012.

Statistics Portugal provided the counts of deaths and population by sex, age group, calendar year and area of residence (parish). A deprivation level was assigned to each parish according to the quintile of their national EDI distribution. Death counts were modelled within the generalised linear model framework, considering a Poisson error with a log link function, using as offset the person-years at risk. Age effect was modelled using restricted cubic splines. Deprivation level, period and interaction between variables were included in the models.

Life expectancy at birth was 74.0 and 80.9 years in $2000 - 2002$, for men and women, respectively, and increased to 77.6 and 83.8 years in 2010-2012. Yet, we observed differences by socioeconomic deprivation: 1.8 and 1.0 years between most and least deprived men and women in 2000-2002. In 2010-2012, the deprivation gap in life expectancy at birth remained similar, at 2.0 and 0.9 years among men and women, respectively. Compared to least deprived, most deprived groups experienced an excess mortality at birth (in 2010-2012, mortality rate ratios of 1.65 and 1.34 in men and women, respectively) which progressively vanished with increasing age.

Substantial and persistent differences in mortality and life expectancy were observed according to area based socioeconomic deprivation. These differences were larger among men and decreased with age for both sexes. No decrease in the deprivation gap was observed between the two periods.

Keywords and key sentences: Life-tables, multivariable modelling, Poisson regression, splines, socioeconomic inequalities in health.

Comunicação poster

Desigualdades sócio-económicas na sobrevivência de doentes oncológicos na presença de informação incompleta

Luís Antunes
Registo Oncológico Regional do Norte, *luis.antunes@ipoporto.min-saude.pt*

Bernard Rachet
London School of Hygiene and Tropical Medicine, *Bernard.Rachet@lshtm.ac.uk*

Fátima Pina
Faculdade de Medicina da Universidade do Porto, *fpina@med.up.pt*

Maria José Bento
Registo Oncolgico Regional do Norte, *mjbento@ipoporto.min-saude.pt*

Denisa Mendonça
Instituto de Ciências Biomédicas Abel Salazar, *dvmendon@icbas.up.pt*

Palavras–chave: sobrevivência, dados omissos

Abstract: A existência de desigualdades sócio-económicas na sobrevivência de doentes com cancro foi já observada em estudos realizados em diferentes países como Inglaterra, França, Estados Unidos ou Austrália. Que seja do nosso conhecimento, nunca foi realizado nenhum estudo para dados portugueses. Neste trabalho pretendeu-se mostrar uma metodologia para avaliar essas desigualdades, ajustando com factores com informação incompleta, aplicando a metodologia a uma amostra de doentes registados no Registo Oncológico Regional do Norte. A condição sócio-económica de cada doente foi obtida com base na respectiva área de residência. A modelação da sobrevivência baseou-se em modelos paramétricos flexíveis e a estimação de valores omissos na utilização de modelos de imputação múltipla.

1 Introdução

Vários estudos mostraram existir associação entre a sobrevivência de doentes oncológicos e a sua condição sócio-económica [1]. Esta associação foi verificada para diferentes localizações topográficas sendo, no entanto, mais evidente em tumores que têm melhor prognóstico e para os quais existe tratamento e possibilidade de cura. Diversos factores podem ajudar a explicar as desigualdades sócio-ecónomicas na sobrevivência, nomeadamente, diferenças no avanço da doença aquando do diagnostico, tratamentos, comorbilidades, entre outros. A informação relativa a estes factores é, no entanto, muitas vezes incompleta, especialmente em registos de base populacional.

2 Métodos

Os registos oncológicos não possuem, habitualmente, informação a um nível individual sobre as condições sócio-económicas dos doentes registados. A atribuição de uma determinada condição a cada doente teve que ser efectuada usando variáveis a nível ecológico, baseada na zona de residência de cada doente. Utilizou-se a informação disponibilizada pelo Instituto Nacional de Estatística relativa aos Censos de 2001. O cruzamento da informação da morada do doente com as regiões geográficas para os quais existe informação censitária foi efectuada utilizando um sistema de informação geográfica. Na modelação da sobrevivência relativa, ajustando com os diferentes factores de prognóstico, foram considerados modelos paramétricos flexíveis [3]. Os valores em falta na extensão da doença, covariável no modelo de sobrevivência, foram estimados usando técnicas de imputação múltipla [4].

3 Aplicação

Pretendeu-se, neste trabalho, avaliar a existência de desigualdades sócio-económicas na sobrevivência de doentes oncológicos. Considerou-se um conjunto de doentes registados no Registo Oncológico Regional do Norte (RORENO). A amostra corresponde a doentes com idade igual ou superior a 15 anos, diagnosticados com tumores malignos do estômago nos anos de 2005 e 2006. A análise incluiu cerca de 590 doentes, dos quais 60% eram do sexo masculino. A extensão da doença encontrava-se omissa em um pouco menos de 20% dos casos. Esta proporção foi maior nos doentes mais velhos e pertencentes ao grupo com tratamento paliativo. Como indicadores sócio-económicos utilizaram-se a proporção de residentes em cada região geográfica com pelo menos a escolaridade obrigatoria e a proporção de desempregados. Resultados preliminares mostraram existir associação entre o nível de escolaridade e a sobrevivência dos doentes, embora a tendência não seja completamente clara. Depois de ajustar para outros factores de prognóstico, como a extensão da doença, sexo e idade, as diferenças entre grupos sócio-económicos foram atenuadas.

Referências

[1] Woods LM, Rachet B and Coleman MP (2005). Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology*, 17(1): 5-19.

[2] Woods LM, Rachet B and Coleman MP (2005). Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. *British Journal of Cancer*, 92(7): 1279-82.

[3] Nelson CP, Lambert PC, Squire IB and Jones, DR (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26(30): 5486-5498.

[4] Royston P, White IR (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*, 45(4).

Poster

Análise de sobrevivência com informação incompleta nas covariáveis - Estudo de simulação

Luís Antunes
Instituto Português de Oncologia do Porto, Instituto de Saúde Pública da Universidade do Porto, luis.antunes@ipoporto.min-saude.pt

Bernard Rachet
London School of Hygiene and Tropical Medicine, Bernard.Rachet@lshtm.ac.uk

Maria José Bento
Instituto Português de Oncologia do Porto, mjbento@ipoporto.min-saude.pt

Denisa Mendonça
Instituto Ciências Biomédicas Abel Salazar, Instituto de Saúde Pública da Universidade do Porto, dvmendon@icbas.up.pt

Palavras–chave: Análise de sobrevivência, Dados omissos, Simulação.

Resumo: A existência de dados omissos em dados na área da saúde, é uma realidade com a qual um bioestatístico se confronta com uma regularidade superior à desejada. No caso concreto de dados de registos oncológicos de base populacional, é frequente encontrar informação em falta em factores de prognóstico importantes como o estadiamento da doença. Esta falta de informação pode levar a que os resultados da análise, que se efectua a esses dados, sejam envieados, especialmente se o mecanismo de omissão não for completamente aleatório.

Pretendeu-se avaliar o desempenho da imputação múltipla como abordagem para lidar com a existência de dados omissos nas covariáveis duma análise de sobrevivência, através de um estudo de simulação, para diferentes proporções de omissão. Foi utilizada como base para o estudo de simulação, uma amostra de dados de sobrevivência correspondente a doentes diagnosticados com tumores gástricos. Os dados foram disponibilizados pelo Registo Oncológico Regional do Norte (RORENO).

Para cada conjunto de dados de sobrevivência simulados, procedeu-se da seguinte forma: eliminação de uma proporção escolhida de casos seguindo padrões de omissão semelhantes aos observados nos dados reais; imputação dos valores omissos usando imputação múltipla; análise de sobrevivência dos dados completados; combinação das diferentes estimativas seguindo as regras de Rubin; comparação dos valores obtidos com os valores reais conhecidos.

A extensão da doença é um dos factores de prognóstico para o qual a proporção de casos omissos é normalmente elevada. O seu valor é completamente definido pelo valor das três variáveis T(tumor), N(nódulos linfáticos) e M(metastização). A ausência do conhecimento da variável T ou da variável N, impede a atribuição do valor da extensão. Observa-se, nos casos reais, que para uma certa proporção de casos apenas uma ou duas destas variáveis se encontra omissa. Utilizando apenas a variável extensão, perde-se a informação disponível nas variáveis T ou N que poderia ser conhecida. Pretendeu-se comparar o comportamento do algoritmo de imputação múltipla em duas situações: imputação directa da variável extensão da doença, sem utilização da informação do TNM e imputação das três variáveis seguida de atribuição do valor da extensão nos dados imputados. A distribuição dos valores imputados das duas formas foi comparada com as distribuições reais, assim como os resultados obtidos no modelo de sobrevivência.

Poster

Desigualdades socioeconómicas na sobrevivência de doentes diagnosticados com tumores do estômago e bexiga na Região Norte de Portugal

Luís Antunes
Instituto Português de Oncologia do Porto, Instituto de Saúde Pública da Universidade do Porto, luis.antunes@ipoporto.min-saude.pt

Bernard Rachet
London School of Hygiene and Tropical Medicine, Bernard.Rachet@lshtm.ac.uk

Maria José Bento
Instituto Português de Oncologia do Porto, mjbento@ipoporto.min-saude.pt

Denisa Mendonça
Instituto Ciências Biomédicas Abel Salazar, Instituto de Saúde Pública da Universidade do Porto, dvmendon@icbas.up.pt

Palavras–chave: Análise de sobrevivência, Factores socioeconómicos, Cancro.

1 Introdução

A análise de sobrevivência de dados de registos de cancro de base populacional é uma importante ferramenta de apoio à decisão. Permite a avaliação dos cuidados de saúde prestados à população coberta por esses mesmos registos e permite a avaliação de heterogeneidades no acesso a esses cuidados. Diferentes estudos têm demonstrado a existência de associação entre as condições socioeconómicas e a sobrevivência de doentes oncológicos. Estas foram já reportadas para países como Inglaterra, Estados Unidos, Austrália, entre muito outros [1]. Para doentes residentes em Portugal não existem, no entanto, resultados publicados sobre esta avaliação.

2 Objectivos

Descrever a sobrevivência de doentes diagnosticados com tumores malignos do estômago ou tumores malignos da bexiga, na Região Norte de Portugal, durante o período 2000-2006. Estudar a associação entre alguns indicadores socioeconómicos e a sobrevivência desses doentes.

3 Material e métodos

Foram incluídos na análise todos os doentes diagnosticados no período de interesse com tumores malignos do estômago ou bexiga, residentes na Região Norte de Portugal e registados pelo Registo Oncológico Regional do Norte (RORENO). A condição socioeconómica de cada doente foi atribuída com base em variáveis ecológicas apenas, visto esta informação não estar disponível a nível individual. O nível geográfico utilizado para esta atribuição foi a freguesia (população mediana: 745). Os indicadores utilizados (nível de escolaridade, analfabetismo, desemprego) foram disponibilizados pelo Instituto Nacional de Estatística e baseam-se na informação obtida nos Censos de 2001 e 2011. Consideraram-se ainda dois indicadores compostos, um indicador de ruralidade e um indicador de acessibilidade a bens e serviços. A sobrevivência relativa foi estimada usando o método Ederer II. O efeito dos factores de prognóstico foi avaliado, estimando Razões de Excesso de Risco (RER), através de modelos paramétricos fléxiveis. Estes permitem uma modelação mais adequada da função de risco de base usando splines cúbicas [2].

4 Resultados

No período de diagnóstico considerado, foram registados 7820 doentes com cancro do estômago e 3630 doentes com cancro da bexiga. A sobrevivência relativa aos 5 anos foi de $33,8\%$ para os tumores do estômago e $73,7\%$ para os tumores da bexiga. A sobrevivência foi significativamente superior nas mulheres em relação aos homens, tanto para os tumores do estômago como para os da bexiga (RER ajustado para a idade: 0.81 e 0.84, respectivamente). Resultados preliminares sugerem que a sobrevivência de doentes residentes em áreas com o menor nível educional e em áreas com o maior índice de ruralidade é significativamente inferior à sobrevivência dos doentes residentes nas restantes áreas.

5 Discussão

Os resultados sugerem que doentes provenientes de áreas mais desfavorecidas apresentam um pior prognóstico, para ambos os tumores analisados. Este pior prognóstico poderá estar relacionado com tendência a diagnósticos da doença em fases mais avançadas. A existência duma grande proporção de informação em falta no estadiamento da doença, não permitiu a validação dessa hipótese. Apesar da dimensão mediana das freguesias ser relativamente baixa, algumas freguesias urbanas apresentam um número de habitantes elevado (acima dos 40 mil), o que poderá ter levado a uma subestimação das desigualdades socioeconómicas na sobrevivência.

Referências

[1] Woods, L.M., Rachet, B., Coleman, M.P. (2006). Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology* 17(1), 5–9.

[2] Nelson, C.P., Lambert, P.C., Squire, I.B., Jones, D.R. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 26(30), 5486–5498.

transverse colon, descending colon, and others were 43.0, 12.5, 10.1, 8.8, 7.9 and 17.7 %, respectively. Stage I, II, III, IV was registered in 6.0, 23.4, 35.3 and 29.5 % cases, respectively, unknown stage in5.8 %. The proportion of urban population (Arkhangelsk, Severodvinsk, Novodvinsk, Kotlas, Koryazhma, Mirny) is 66.4 % of all diagnosed cases of CC in the Arkhangelsk region. During the period the standardized incidence rates of CC have increased from 14.0 to 17.8 per 100,000. The crude incidence among men and women have increased from13.6 per 100,000 in 2000 to 22.1 per 100,000 in 2010 and from 21.0 per 100,000 in 2000 to 28.8 in 2010, respectively. The incidence among the urban population has risen from 15.5 per 100,000 in 2000 to 24.4 per 100,000 in 2010. The incidence in rural areas was slightly higher and ranged from 23.3 per 100,000 in 2000 to 29.2 per 100,000 in 2010.

Conclusions: The incidence and the mortality of CC in the Arkhangelsk region of Russia increased during the period 2000–2010, resembling incidence pattern in countries in transition. Most of the patients are females. The incidence is higher among rural population.

P-089

Impact of diabetes mellitus on long-term survival of hepatocellular carcinoma

Topic: Diseases - Cancer
Presenter: Elena Raffetti, Francesco Donato

Elena Raffetti 1); Giovanni Caccamo 1); Rossella Lamera 1); Sarah Molfino 2); Andrea Celotti 2); Rosa Maria Limina 1); Arianna Coniglio 2); Nazario Portolani 2); Francesco Donato 1)

1) Unit of Hygiene, Epidemiology and Public Health, University of Brescia, Italy; 2) Surgical Clinic, University of Brescia, Italy

Background: The influence of diabetes mellitus (DM) on hepatocellular carcinoma (HCC) incidence remains obscure and it is not clear whether it may affect the overall survival.

Objective: Evaluation of DM influence on HCC survival.

Methods: We prospectively enrolled 329 patients, with first diagnosis of HCC from 1995 to 2001, in Brescia, Italy, Etiology was assessed by interviewing patients regarding their history of alcohol intake and by testing sera for hepatitis B surface antigen and anti-hepatitis C virus (HCV) antibodies and HCV RNA. Patients was considered to be diabetics in presence of hospital discharge DM code. Survival was determined from the date of HCC diagnosis to the end of follow-up, which was December 31, 2012. Cumulative survival curves were modeled by using the Kaplan-Meier method. The association of each variable with patient survival was tested by univariate analysis using the log-rank test. The same variables were tested by multivariate analysis using Cox proportional hazard models.

Results: Among 329 patients with HCC 271 (82.4 %) were males and 98 (29.8 %) had DM. Heavy alcohol intake (>40 g in men and >20 g in women of ethanol per day for at least 1 decade) was found in 36.5 % of cases, hepatitis virus infection in 15.5 %, alcohol and hepatitis virus infection in 41.9 % and other factor in 6.1 %. Thirteen patients (4.0 %) were alive at the end of follow-up, with a median survival of 19.8 months (IC95 % 17.1–22.5). Overall survival at 1, 5 and 10 years was 61.3, 22.9 and 7.6 % respectively. On multivariate analysis, survival was associated with serum ALT > 100 U/l (hazard ratio [HR] = 1.4, $p = 0.018$), stage pT (HR = 1.2; $p = 0.023$), portal vein invasion (HR = 1.9; $p < 0.001$), cirrhosis (HR = 1.66, $p = 0.001$), metastasis (HR = 2.7, $p = 0.001$), treatment (radical treatment vs. palliative treatments/no treatment HR = 3.0, $p < 0.001$), Child classification (A vs. B HR = 1.38, $p = 0.032$; A vs. C HR = 1.82, $p = 0.003$). The presence of DM was negatively associated with survival (HR 1.3, $p = 0.07$).

Conclusions: In this study a higher prevalence of DM was found in HCC patients as compared to italian general population in the same age group (12.5 %) and negative influence of DM on HCC survival was observed.

P-090

Socioeconomic inequalities in cancer survival for the most common cancer sites in the North Region of Portugal, 2000–2006

Topic: Diseases - Cancer
Presenter: Denisa Mendonça

Luis Antunes 1); Bernard Rachet 2), 3); Maria Jose Bento 1); Denisa Mendonça 4), 5)

1) Registo Oncológico Regional do Norte (RORENO), Instituto Português de Oncologia do Porto; 2) Cancer Survival Group, London School of Hygiene and Tropical Medicine (LSHTM); 3) Centre for Cancer Control and Statistics, Osaka Medical Centre for Cancer and Cardiovascular diseases (OMCC); 4) Instituto Ciências Biomédicas Abel Salazar (ICBAS); 5) Instituto Saúde Pública da Universidade do Porto (ISPUP)

Background: Socioeconomic inequalities in cancer survival have been consistently reported in different countries for most of the adult cancers. However, this has not yet been confirmed for Portuguese cancer patients.

Objectives: To describe the survival of patients diagnosed in the North Region of Portugal with one of the most common cancers (stomach, colorectum, lung, breast, bladder) during the period 2000–2006 and to study the influence of socioeconomic conditions in survival.

Materials and methods: Data consisted in cancer patients resident in the North Region of Portugal and registered by the corresponding population-based cancer registry (RORENO). All malignant, invasive, primary tumours of breast, colorectum, stomach, lung and bladder diagnosed among adult in 2000–2006, were considered for analysis. Socioeconomic conditions were assigned to each patient using ecological variables defined at parish level, namely, level of education, illiteracy, unemployment, index of rurality and index of accessibility to goods and services. The levels for each geographical area and each year of diagnosis were estimated based on the information of two different population census (2001 and 2011). Relative survival was calculated using Ederer II method. Excess hazards ratios (EHR) were estimated using a flexible modelling approach enabling to model the effects of sex, age and socio-economic condition.

Results: A total of 40,768 patients were diagnosed in the period of interest (breast-27 %; colorectal-29 %; stomach-19 %; lung-16 %; bladder-9 %). Five-year relative survival was lower for lung (10.5 %) and stomach cancer (33.8 %) and higher for colorectal (59.7 %), bladder (73.7 %) and breast cancer (87.0 %). For stomach, lung and bladder cancer, women had a better survival than men (EHR adjusted for age: 0.81, 0.79 and 0.84, respectively) while for colorectal, no differences were found EHR = 0.98). Preliminary results have shown that stomach and bladder cancer patients coming from areas with a higher index of rurality or lower level of education have lower survival, while for colorectal, lung and breast cancers the survival rates are similar across socioeconomic levels.

Discussion: Socioeconomic inequalities in cancer survival were found more significant for bladder and stomach cancers. The median number of individuals by parish is relatively small (745) but some urban areas reach more than forty thousand inhabitants, what can lead to an underestimation of the socioeconomic gap in survival. The proportion of missing information on stage of disease at diagnosis was higher than fifty percent, precluding the inclusion of this variable in the analysis.

COLORECTAL CANCER SURVIVAL BY EDUCATION LEVEL IN AN URBAN AREA OF THE NORTH REGION OF PORTUGAL, 2000-2002

Luís Antunes[1]; Maria José Bento[1]; Clara Castro[1]; Bernard Rachet[2]; Denisa Mendonça[3,4]

[1] Registo Oncológico Regional do Norte (RORENO), Instituto Português de Oncologia do Porto
[2] Cancer Survival Group, London School of Hygiene and Tropical Medicine (LSHTM)
[3] Instituto Ciências Biomédicas Abel Salazar (ICBAS)
[4] Instituto Saúde Pública da Universidade do Porto (ISPUP)

e-mail: luis.antunes@ipoporto.min-saude.pt

Introduction

Socioeconomic conditions are known to affect cancer survival. Many factors can contribute to these inequalities, including differential access to diagnosis/treatment centres and different cancer symptoms awareness.

Objectives

To understand the role of education level as a prognostic factor for colorectal cancer patient's survival using a population-based dataset from a large urban area where geographical access to health care centres is homogeneous.

Material and Methods

All malignant invasive colorectal cancer patients (ICD-10: C18-C20), with residence at diagnosis in the city of Porto, diagnosed in the period 2000 to 2002, aged 15 years or older, were considered for analysis. Education level was assigned to each patient based on the area of residence at census tract level and measured by the proportion of residents with at least the compulsory level of education. Net survival was estimated using Pohar-Perme estimator and age-adjusted excess hazards (EHR) were estimated using parametric flexible models.

Results

A total of 550 patients (51.5% male) were considered eligible for analysis. After excluding cases with no follow-up information (1.1%), 544 cases were included for analysis. Overall 5-year net survival was 58.5% (95%CI: 53.4-63.5). No differences in survival were found by sex (p=0.312). Patients were grouped by education level of its area of residence in three groups: very low education, medium education and very high education. Five-year net survival ranged from 53.9% (95%CI: 40.9-67.0) for the lower educated group to 72.5% (95%CI: 60.7-84.2) for the highest group.

The excess hazard in the more educated patients was lower, although not reaching statistical significance (age-adjusted EHR: 0.64; 95%CI: 0.36-1.13) while the excess hazard in the medium educated group was similar to the one of the less educated group (age-adjusted EHR: 0.95; 95%CI: 0.63-1.44).

Discussion and conclusions

A higher survival for the group of patients coming from the highest educated areas has been observed, although not statistically significant probably due to small number of cases. By considering as region of interest an urban area, the differences in survival by education level may be more likely attributable to other causes than differences in geographical access to treatment centres, namely, different cancer symptoms awareness or comorbidities.

DESIGUALDADES SOCIOECONÓMICAS NA SOBREVIVÊNCIA DE DOENTES COM CANCRO COLO-RECTAL NA CIDADE DO PORTO

Luís Antunes[1]; Maria Fátima Pina[2,3,5]; Maria José Bento[1]; Denisa Mendonça[4,5]

[1] Registo Oncológico Regional do Norte (RORENO), Instituto Português de Oncologia do Porto
[2] Faculdade de Medicina da Universidade do Porto (FMUP)
[3] Instituto Nacional de Engenharia Biomédia (INEB)
[4] Instituto Ciências Biomédicas Abel Salazar (ICBAS)
[5] Instituto Saúde Pública da Universidade do Porto (ISPUP)

e-mail: luis.antunes@ipoporto.min-saude.pt

Antecedentes/Objectivos

A condição socioeconómica de um doente oncológico é um reconhecido factor de prognóstico. Vários factores podem contribuir para estas iniquidades, incluindo heterogeneidade no acesso a centros de diagnóstico e tratamento, diferentes comorbilidades ou diferente valorização dos sinais e sintomas da doença. Pretendeu-se neste trabalho, avaliar a sobrevivência por nível de privação socioeconómica de doentes residentes na cidade do Porto, diagnosticados com cancro colo-rectal.

Métodos

Foram considerados elegíveis, todos os doentes diagnosticados no período 2000-2002, com tumores colo-rectais (IDC10: C18-C20), com residência no Porto e idade igual ou superior a 15 anos. A condição socioeconómica de cada doente foi atribuída com base na área de residência, ao nível da subsecção estatística. O indicador utilizado agrega informação relativa à distribuição etária, educação, ocupação e condição das habitações. A sobrevivência net foi estimada usando o estimador de Pohar-Perme e as razões de excesso de risco foram estimadas usando modelos paramétricos flexíveis.

Resultados

Foram identificados 550 doentes elegíveis para análise (51,5% do sexo masculino). Após exclusão de casos sem informação de follow-up ou sem informação da condição socioeconómica (2,5%), foram considerados 536 casos. A sobrevivência net aos 5 anos variou nos doentes do sexo masculino entre 64,3% e 60,9% (nos grupo mais e menos favorecido, respectivamente). Para as doentes do sexo feminino, a sobrevivência net aos 5 anos variou entre 72,7% no grupo mais favorecido e 44,8% no grupo menos favorecido. O grupo com maior índice de privação apresentou um excesso de risco de morte (ajustado para a idade) significativamente superior ao do grupo mais favorecido (RER=2,25; IC95: 1,18-4,27) nas mulheres, enquanto para os homens, o excesso de risco nos dois grupos foi semelhante (RER=1,03; IC95: 0,53-2,01).

Conclusões

Observaram-se desigualdades na sobrevivência por grupo socioeconómico nos doentes do sexo feminino mas não nos do sexo masculino. Dado que a análise se restringiu a um meio urbano, as desigualdades atribuíveis a diferenças na acessibilidade geográfica aos centros de diagnóstico/tratamento deverão ser mínimas. Outras causas possíveis poderão estar relacionadas com maior atraso na procura de cuidados médicos por parte das doentes com maior índice de privação, no entanto, as razões para estas desigualdades necessitam de ser investigadas com maior detalhe.

Abstract #: P 01

Identification and hierarquization the risk factors for colorectal cancer in Alentejo Litoral: a case control epidemiological study

Sara Letras[1], Pedro Aguiar[2], Mário Jorge Santos[3]

[1]Public Health Unit, Local Health Unit of Alentejo Litoral (ULSLA); [2]National School of Public Health, Universidade Nova de Lisboa; [3]Public Health Services, Local Health Unit of Alentejo Litoral (ULSLA); Corresponding author's e-mail: saraletras@gmail.com

Background: Colorectal cancer (CRC) is currently considered a major public health problem internationally and for Portugal. Due to population aging in Alentejo Litoral (AL), the problem get a major magnitude (annual average incidence being around 80 new cases/100,000 inhabitants, annual prevalence rate up to 500 cases/100,000, average annual mortality up to 40 deaths/100,000) and that reinforces the importance to identify the main risk factors (RF) for CRC in order to address possible and to maximize effective prevention measures. The main goal of the study it was the identification and its ranking for the main risk factors to CRC.

Methods: It was performed a retrospective analytical epidemiological case–control study, with the Odds ratio (OR) determination. The observation unit it was: to be resident in the AL, and have greater than or equal to 40 years. The study included 90 CRC cases (identified by ROR-Sul) and 201 controls (general medical consultations). Cases and controls were inquired for independent variables included in the study (sociodemographic, family/personal and behavioral history). Descriptive, bivariate and multivariate (logistic regression) statistical analysis (SPPS 20) was performed.

Results: The main results for bivariate analysis (Confidence Interval (CI) 95 % and p value < 0.10) were: personal history of inflammatory bowel disease (OR = 9.302): Insulin therapy previous to CRC (OR = 6.897): consumption of alcoholic beverages in the past (OR = 4.853) and with a typical frequency greater than 4 times a week (OR = 3.632). The main results for multivariate analysis (CI 95 % and $p < 0.05$) were: consumption of red meat (adjusted OR = 6.828), family history for CRC (adjusted OR = 6.628) and number of alcoholic drinks (in a typical day) greater than or equal to 3 per day (adjusted OR = 5808).

Conclusion: The main RF for CRC in AL were: red meat consupmtion, family history for CRC and number of alcoholic drinks greater than or equal to 3 per day. The study also concludes that most RF identified for this population are modifiable, can be targeted for interventions of health promotion and disease prevention and establish wich those RF are more cost effective regarding possible health interventions.

Abstract #: P 02

Socioeconomic position and incidence of colorectal cancer in the Swedish population

Hannah L. Brooke[1], Mats Talbäck[1], Anna Martling[2], Maria Feychting[1], Rickard Ljung[1]

[1]Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; [2]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; Corresponding author's e-mail: hannah.brooke@ki.se

Background: The association between socioeconomic position and incidence of colorectal cancer is unclear. We aimed to clarify this association, in the whole Swedish population. This work may inform policy and prevention strategies designed to reduce health inequalities.

Methods: We conducted a population-based open cohort study using national registry data. We included all individuals, aged ≥30 years, residing in Sweden between 1993 and 2010, without a previous diagnosis of colon or rectal cancer. Socioeconomic position was indicated by (1) highest education level ('≤primary', 'lower secondary', 'higher secondary', 'lower university' [<3 years], 'higher university' [≥3 years]), and (2) personal disposable income (quintiles). The outcome was diagnosis of colon or rectal cancer. We used Poisson regression to estimate incidence rate ratios (IRR) and 95 % confidence intervals (95 % CI) of colon and rectal cancer, for each exposure. Models were stratified by sex and adjusted for age, year of follow-up, region of residence, and marital status, with mutual adjustment of exposures.

Results: In 100,679,466 person-years of follow-up, 61,793 cases of colon cancer (30,014 men, 31,779 women) and 30,131 cases of rectal cancer (17,379 men, 12,752 women) were diagnosed.

In men and women, IRRs of colon cancer were close to 1.00 for all education levels compared with the least educated, and for all quintiles of personalised disposable income compared with the lowest quintile. However, there was a slightly higher risk of colon cancer in men with 'higher secondary' compared with '≤primary' education (IRR [95 % CI]: 1.05 [1.02, 1.09]). In women, there was a higher risk of colon cancer in the middle compared with the lowest quintile of personal disposable income (IRR [95 % CI]: 1.07 [1.03, 1.11]).

Risk of rectal cancer in men and women gradually decreased with increasing education level. Compared with '≤primary' education, the IRRs (95 % CI) of rectal cancer in men with 'lower secondary', 'higher secondary', 'lower university' or 'higher university' education were: 0.99 (0.95, 1.03), 0.94 (0.90, 0.99), 0.90 (0.84, 0.96), and 0.86 (0.81, 0.92), respectively. In women, the corresponding figures were: 0.99 (0.95, 1.04), 0.95 (0.87, 1.04), 0.88 (0.82, 0.96) and 0.87 (0.81, 0.94). Rectal cancer incidence did not differ between quintiles of personal disposable income.

Conclusions: In the Swedish population, incidence of colon cancer was not clearly patterned by socioeconomic position. However, there was lower incidence of rectal cancer in more highly educated groups. To help reduce such health inequalities, further evaluation of potentially preventable mechanisms and health promotion strategies among deprived groups is warranted.

Abstract #: P 03

Impact of the choice of life tables on the assessment of socioeconomic inequalities in survival from colorectal cancer

Luis Antunes[1], Bernard Rachet[2], Maria José Bento[1], Denisa Mendonça[3]

[1]North Region Cancer Registry of Portugal, Porto, Portugal; [2]Cancer Survival Group, London School of Hygiene and Tropical Medicine, London, UK; [3]EPIUnit - Institute of Public Health, University of Porto, Porto, Portugal; Corresponding author's e-mail: luis.antunes@ipoporto.min-saude.pt

Background: Socioeconomic condition can affect the mortality of a cancer patient from both their cancer and other causes. Assessing socioeconomic inequalities in cancer survival must therefore account for socioeconomic inequalities in background mortality. When the cause of death is unknown (relative survival data setting), net survival is estimated by comparing observed survival with background mortality provided by general population life tables (LTs).

We aimed at evaluating the impact of the choice of LTs in estimating inequalities in colorectal cancer survival by education group.

Methods: All patients, aged 15–84, diagnosed with a malignant invasive colorectal cancer (ICD-10: C18–C20), in the North Region of Portugal in 2000–2002, were considered for analysis. Education level, the proportion of residents with at least the compulsory level of education in a given census tract, was assigned to each patient based on their area of residence and categorised according to quintiles. Net survival and age-adjusted excess hazards ratios (EHR) were estimated using Pohar-Perme estimator and flexible parametric models, respectively. Three different scenarios were considered for background mortality: no education-specific Portuguese LTs (S1): education-specific LTs considering the same ratios between socioeconomic groups as observed in England (S2): education-specific LTs with a 70 % reduction (relative to S2) in the log ratio between socioeconomic groups (S3).

Results: A total of 4105 patients (56.3 % male) were analysed. In scenario S1, male patients coming from lower educated areas had comparable 1-year survival to patients from higher educated areas, but a lower 5- and 10-year survival (at 5 years: EHR = 1.40, 95 % CI 1.07–1.83: at 10 years: EHR = 1.56, 95 % CI 1.08–2.25). Inequalities in survival decreased substantially in both scenarios S2 (5 years: EHR = 1.09, 95 % CI 0.83–1.43: 10-yrs: EHR = 1.12, 95 % CI 0.77–1.61) and S3 (5 years: EHR = 1.30, 95 % CI 0.99–1.70: 10 years: EHR = 1.41, 95 % CI 0.98–2.03).

No significant differences in survival were found in women.

Conclusion: No education-specific life tables are available for Portugal. To test the sensitivity of the inequalities found in men to the choice of the LT, we built two sets of education-specific LT in which differences in life expectancy at birth between extreme groups were 7.7 years (S2) and 2.3 years (S3). Cancer survival inequalities observed in S1 faded out in both scenarios, suggesting that the observed differences are most likely attributable to education inequalities in background mortality, stressing the importance of using the adequate life tables in cancer survival inequalities assessment.

Abstract #: P 04

Changes in body weight during and after treatment for colorectal cancer

Renate M. Winkels[1], Teunise Snetselaar[1], Anika Adriaans[2], Laurence J.C. van Warmerdam[3], Art Vreugdenhil[4], Gerrit Slooter[4], Jan-Willem Straathof[4], Ellen Kampman[1], Rianne van Lieshout[4], Sandra Beijer[2]

[1]Agrotechnology and Food Sciences, Wageningen University, Wageningen, the Netherlands; [2]Netherlands Cancer Registry, Eindhoven, the Netherlands; [3]Catharina Hospital, Eindhoven, the Netherlands; [4]Máxima Medical Centre, Eindhoven, the Netherlands; Corresponding author's e-mail: teunise_s@hotmail.com

Background: Prevalence of overweight and obesity is high among colorectal cancer patients at diagnosis. Literature suggests that body weight may further increase during adjuvant chemotherapy for colorectal cancer. However, so far, weight changes from diagnosis until after treatment have not been studied in this patient group.

Methods: The study population consisted of 485 stage II/III colorectal cancer patients diagnosed between 2007 and 2012 and treated with surgery and adjuvant chemotherapy in one of three selected hospitals in the Netherlands. Eligible patients were selected from the Netherlands Cancer Registry. Data about body weight (at diagnosis, shortly after surgery, shortly after chemotherapy and during follow-up) and other personal/clinical factors were retrieved from the cancer registry and from medical records.

Results: From diagnosis until shortly after surgery, patients on average lost weight (mean weight loss −1.9 kg, SD 4.6 kg) (n = 357). Body weight increased during chemotherapy with a mean of 2.9 kg (SD 5.8 kg) (n = 291) and continued to increase in the period of follow-up by 2.2 kg (SD 6.6 kg) (n = 242). Overall, from diagnosis until at least 6 months after chemotherapy, there was a mean weight gain of 2.0 kg (SD 6.8 kg) (n = 283). Factors associated with weight gain over this period were a normal BMI (vs patients with a BMI of 25–30), open surgery (vs laparoscopic surgery) and Capecitabine chemotherapy (vs Capecitabine in combination with Oxaliplatin).

Conclusions: Body weight generally decreased from diagnosis until shortly after surgery, while it increased again during and after chemotherapy. At least 6 months after chemotherapy, body weight was higher than at diagnosis. Studies among other patient groups—mostly breast cancer—suggest that these changes may be characterised by unbeneficial changes in body composition, e.g. sarcopenic obesity. Future studies should characterize changes in body weight and composition and the impact on the health and quality of life of colorectal cancer patients.

Abstract #: P 05

Interlaboratory variability in grading of dysplasia in a nationwide cohort of colorectal adenomas

Chantal C.H.J. Kuijpers[1,2,3], Caro E. Sluijter[2,4], Lucy I.H. Overbeek[2], Jan H. von der Thüsen[5,6], Katrien Grünberg[6,7], Paul J. van Diest[1], Mehdi Jiwa[1,3], Iris D. Nagtegaal[2,5], Stefan M. Willems[1,2]

[1]Department of Pathology, University Medical Centre Utrecht, Utrecht, The Netherlands; [2]Foundation PALGA, Houten, The Netherlands; [3]Symbiant Pathology Expert Centre, Alkmaar, The Netherlands; [4]Department of Pathology, Radboud University Medical Centre, Nijmegen, The Netherlands; [5]Department of Pathology, Medical Centre Haaglanden, The Hague, The Netherlands; [6]NVVP (Dutch Society of Pathology), Utrecht, The Netherlands; [7]Department of Pathology, VU University Medical Centre, Amsterdam, The Netherlands; Corresponding author's e-mail: c.c.h.kuijpers@umcutrecht.nl

Background: Colorectal adenomas are precursor lesions of colorectal adenocarcinoma. One of the risk factors for malignant transformation and future development of a new adenoma or carcinoma is the presence of high-grade dysplasia (HGD). However, this factor is not incorporated in Dutch colonoscopy surveillance guidelines, partly due to high variability in grading dysplasia between pathologists. We aimed to determine, on a nationwide basis, whether histological grading of colorectal adenomas varies in daily practice between Dutch pathology laboratories.

Methods: Using the Dutch Pathology Registry (PALGA), all synoptic pathology reports of colorectal biopsies and polypectomies histologically diagnosed in 2013 as tubular, tubulovillous or villous adenoma were identified. Percentages of low-grade dysplasia (LGD) and HGD were determined for biopsies and polypectomies separately, and clinico-pathological factors associated with HGD were investigated. In a subgroup of 21 Dutch pathology laboratories, each with $\geq$100 synoptically reported colorectal adenomas, percentages of HGD per laboratory were compared. Univariable and multivariable logistic regression analyses were performed.

Results: Pathology reports of 21,145 colonoscopies of 20,332 patients (57 % males, mean age: 66 year) with $\geq$1 adenomas were identified. The 32,524 histologically confirmed adenomas included 21,544 adenomas from biopsies and 10,980 adenomas from polypectomies. HGD was diagnosed in 2.6 % and 9.6 % of adenomas from biopsies and polypectomies, respectively. In both subgroups, HGD was significantly associated with advanced age, distal location,

Estimation of age-standardized net survival with sparse data: taking advantage of regression models

Luís Antunes
Faculdade de Ciências da Universidade do Porto; Instituto Português de Oncologia do Porto, *luis.antunes@ipoporto.min-saude.pt*

Denisa Mendonça
Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, *dvmendon@icbas.up.pt*

Aurelien Belot
Cancer Survival Group, London School of Hygiene and Tropical Medicine, *Aurelien.Belot@lshtm.ac.uk*

Bernard Rachet
Cancer Survival Group, London School of Hygiene and Tropical Medicine, *Bernard.Rachet@lshtm.ac.uk*

Palavras–chave: net survival, age standardisation, sparse data

Abstract:

Cancer survival analysis is of major importance in the evaluation of cancer care practices provided to populations. International comparison of survival probabilities from cancer should take into account differences in patientŠs population age structure since survival from cancer is often age dependent. This is usually achieved through direct age-standardization using a common age distribution standard such as the International Cancer Survival Standards. The direct age-standardization implies the estimation of survival for each age group. Often, the extreme age groups (youngest or oldest, depending on the cancer) are sparse and their net survival estimates are either very unstable or even impossible to obtain a few years after diagnosis.

Net survival, the survival that would be observed in the absence of causes of death not related to the disease in study, can be estimated using the Pohar-Perme estimator or a modelling approach. If the model is correctly specified, both methods should produce the same estimate. When age is considered as a continuous variable and the excess hazard is modelled with flexible functions (e.g. splines), net survival of each individual can be thinly predicted for any time since diagnosis. The net survival of a given age group is obtained as the mean of the individual net survival of the subjects in this age group. Although a flexible modelling approach is used, net survival estimate of each age group

depends on the observed number of subjects in each group as well as on their observed age-distribution. This will again lead to unstable net survival estimates when the data are sparse even if the model allows to smoothly predict exact individual net survivals. Age group-specific estimates given by the non-parametric Pohar-Perme estimator are also very unstable on such datasets.

The main aim of this study was to evaluate and compare methods for the estimation of age-standardized net survival when data are sparse, using a simulation study. Different approaches were compared: model-based predictions and non-parametric estimates.

Three different scenarios were considered with increasing model complexity. Large datasets ($N = 10^6$) using models fitted to real cancer datasets were generated for each scenario. From these sets, we randomly selected 1000 small samples ($n = 200$). For each sample, four model fitting approaches were used: same type of model for all samples (non-linear and time-dependent effects of age); choose "best" model for each sample; categorical age; semi-continuous age (categorized only in extreme age groups). Model-based net survival predictions were obtained averaging individual predictions (established approach) and predicting for a reference age in each age group (alternative proposed approach). Additionally, net survival was estimated using the Pohar-Perme estimator. Net survival was age-standardised by weight averaging age-group specific estimates. Estimates were produced for the full samples and for smaller subsets.

The estimates obtained using the established and alternative approaches had similar performance in terms of bias and coverage probability when estimating survival for the full samples. However, when estimating for smaller subsets, the alternative approach allowed the estimation of survival for a much higher proportion of samples than the classical approach.

These results suggest that, for situations where data are sparse, an alternative estimation approach could be used. Further studies with more complex scenarios are under way to confirm, or not, the feasibility of these alternative approaches.

Appendix B

R Code

In this appendix, the R code developed for some of scenarios tested in the simulation studies and statistical analysis performed along this thesis is presented. For the other different scenarios tested the code is similar and is not presented.
More specifically, the code for the following tasks is presented:

- Evaluation of age-standardise net survival estimators (non-parametric and model-based). Simulation study for scenario B (Non-linear and time-dependent efect of age, linear and proportional efect of year of diagnosis).

- Simulation study to evaluate performance of SMC-FCS algorithm for excess hazard models comparing with complete case analysis and standard FCS multiple imputation. Code for scenario C (outcome-dependent MAR).

- Evaluation of socioeconomic inequalities in survival from colorectal cancer. Comparison of complete-case analysis, standard FCS multiple imputation and SMC-FCS. Code for SMC-FCS when substantive model is an excess hazard model and missingness in categorical covariate.

```
###########################################################################
# Age-standardised net survival estimation
#
# Comparison of model-based approaches and PP
#
# Scenario B
#
# Breast cancer
# Choose best model for each sample
#
# Simulated data files were generated elsewhere
#
##########################################################################
library(mexhaz)
library(relsurv)
library(statmod)

setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo1/Paper_simulations")
source("C:\\Users\\ljant\\Documents\\Doutoramento - PDMA\\Artigo1\\Paper_Simulations\\NewSurvPop_Hadrien.r")

set.seed(123456)

# Sample size
n_size=2000

# Number of simulations by cycle
n_sim=1000

# Weights for age standardization
weights=c(0.07,0.12,0.23,0.29,0.29)

# To save survival by age profile for each chosen model
# age vector
mean_age=60
sd_age=14
knot=60
age_vector=seq(15,99,1)
n_age=length(age_vector)
agecrk=rep((knot-mean_age)/sd_age,n_age)
agecr_vector=(age_vector-mean_age)/sd_age
agecr_vector2=agecr_vector^2
agecr_vector3=agecr_vector^3
agecr_vectortr=as.numeric(age_vector>knot)*(agecr_vector-agecrk)^3
agecr_yydx_vector=0

# Survival matrices
surv_agegrp_MB1=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB1_2001=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB1_2010=matrix(NA,nrow=n_sim,ncol=5)

surv_agegrp_MB2=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2001=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2002=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2003=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2004=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2005=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2006=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2007=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2008=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2009=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_MB2_2010=matrix(NA,nrow=n_sim,ncol=5)

# Standard error matrices
stand_agegrp_MB1=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB1_2001=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB1_2010=matrix(NA,nrow=n_sim,ncol=5)

stand_agegrp_MB2=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2001=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2002=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2003=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2004=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2005=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2006=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2007=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2008=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2009=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_MB2_2010=matrix(NA,nrow=n_sim,ncol=5)

# Age-standardised survivals
asns_MB1=asns_MB1_2001=asns_MB1_2010=NULL
se_MB1=se_MB1_2001=se_MB1_2010=NULL
asns_MB2=matrix(NA,nrow=n_sim,ncol=10)
se_MB2=matrix(NA,nrow=n_sim,ncol=10)

asns_MB2_ave=NULL
se_MB2_ave=NULL
```

```
# Survival profiles by age
surv_by_age=matrix(NA,nrow=n_sim,ncol=n_age)

# Get reference age for MB2
m1_agec=read.table(file="scB_ref_agec.txt",sep="\t")
m1_agec2=read.table(file="scB_ref_sec2.txt",sep="\t")
m1_agec3=read.table(file="scB_ref_agec3.txt",sep="\t")
m1_ageknot=read.table(file="scB_ref_ageknot.txt",sep="\t")

m1_agec=m1_agec[,1]
m1_agec2=m1_agec2[,1]
m1_agec3=m1_agec3[,1]
m1_ageknot=m1_ageknot[,1]

# Fake populations for MB2
mydata_mb2_2001=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2002=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2003=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2004=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2005=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2006=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2007=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2008=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2009=data.frame(NA,nrow=5,ncol=6)
mydata_mb2_2010=data.frame(NA,nrow=5,ncol=6)

mydata_mb2=data.frame(NA,nrow=50,ncol=6)

for(j in 1:5) {
mydata_mb2_2001[j,1]=m1_agec[j]
mydata_mb2_2002[j,1]=m1_agec[j]
mydata_mb2_2003[j,1]=m1_agec[j]
  mydata_mb2_2004[j,1]=m1_agec[j]
  mydata_mb2_2005[j,1]=m1_agec[j]
  mydata_mb2_2006[j,1]=m1_agec[j]
  mydata_mb2_2007[j,1]=m1_agec[j]
  mydata_mb2_2008[j,1]=m1_agec[j]
  mydata_mb2_2009[j,1]=m1_agec[j]
  mydata_mb2_2010[j,1]=m1_agec[j]

  mydata_mb2_2001[j,2]=m1_agec2[j]
  mydata_mb2_2002[j,2]=m1_agec2[j]
  mydata_mb2_2003[j,2]=m1_agec2[j]
  mydata_mb2_2004[j,2]=m1_agec2[j]
  mydata_mb2_2005[j,2]=m1_agec2[j]
  mydata_mb2_2006[j,2]=m1_agec2[j]
  mydata_mb2_2007[j,2]=m1_agec2[j]
  mydata_mb2_2008[j,2]=m1_agec2[j]
  mydata_mb2_2009[j,2]=m1_agec2[j]
  mydata_mb2_2010[j,2]=m1_agec2[j]

  mydata_mb2_2001[j,3]=m1_agec3[j]
  mydata_mb2_2002[j,3]=m1_agec3[j]
  mydata_mb2_2003[j,3]=m1_agec3[j]
  mydata_mb2_2004[j,3]=m1_agec3[j]
  mydata_mb2_2005[j,3]=m1_agec3[j]
  mydata_mb2_2006[j,3]=m1_agec3[j]
  mydata_mb2_2007[j,3]=m1_agec3[j]
  mydata_mb2_2008[j,3]=m1_agec3[j]
  mydata_mb2_2009[j,3]=m1_agec3[j]
  mydata_mb2_2010[j,3]=m1_agec3[j]

  mydata_mb2_2001[j,4]=m1_ageknot[j]
  mydata_mb2_2002[j,4]=m1_ageknot[j]
  mydata_mb2_2003[j,4]=m1_ageknot[j]
  mydata_mb2_2004[j,4]=m1_ageknot[j]
  mydata_mb2_2005[j,4]=m1_ageknot[j]
  mydata_mb2_2006[j,4]=m1_ageknot[j]
  mydata_mb2_2007[j,4]=m1_ageknot[j]
  mydata_mb2_2008[j,4]=m1_ageknot[j]
  mydata_mb2_2009[j,4]=m1_ageknot[j]
  mydata_mb2_2010[j,4]=m1_ageknot[j]

  mydata_mb2_2001[j,5]=weights[j]
  mydata_mb2_2002[j,5]=weights[j]
  mydata_mb2_2003[j,5]=weights[j]
  mydata_mb2_2004[j,5]=weights[j]
  mydata_mb2_2005[j,5]=weights[j]
  mydata_mb2_2006[j,5]=weights[j]
  mydata_mb2_2007[j,5]=weights[j]
  mydata_mb2_2008[j,5]=weights[j]
  mydata_mb2_2009[j,5]=weights[j]
  mydata_mb2_2010[j,5]=weights[j]

  mydata_mb2_2001[j,6]=2001-2005
  mydata_mb2_2002[j,6]=2002-2005
  mydata_mb2_2003[j,6]=2003-2005
  mydata_mb2_2004[j,6]=2004-2005
  mydata_mb2_2005[j,6]=2005-2005
  mydata_mb2_2006[j,6]=2006-2005
  mydata_mb2_2007[j,6]=2007-2005
  mydata_mb2_2008[j,6]=2008-2005
  mydata_mb2_2009[j,6]=2009-2005
  mydata_mb2_2010[j,6]=2010-2005

  mydata_mb2_2001[j,7]=m1_agec[j]*(2001-2005)
  mydata_mb2_2002[j,7]=m1_agec[j]*(2002-2005)
  mydata_mb2_2003[j,7]=m1_agec[j]*(2003-2005)
  mydata_mb2_2004[j,7]=m1_agec[j]*(2004-2005)
  mydata_mb2_2005[j,7]=m1_agec[j]*(2005-2005)
  mydata_mb2_2006[j,7]=m1_agec[j]*(2006-2005)
  mydata_mb2_2007[j,7]=m1_agec[j]*(2007-2005)
  mydata_mb2_2008[j,7]=m1_agec[j]*(2008-2005)
  mydata_mb2_2009[j,7]=m1_agec[j]*(2009-2005)
  mydata_mb2_2010[j,7]=m1_agec[j]*(2010-2005)
}

colnames(mydata_mb2_2001)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2002)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2003)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2004)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2005)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2006)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2007)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2008)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2009)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")
colnames(mydata_mb2_2010)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")

setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo1/UM_SeminarioFev2017/Samples/2000/B")

count_agegrp=read.table(file=paste(n_size,"scB_count_agegrp.txt",sep=""),sep="\t")
count_agegrp_2001=read.table(file=paste(n_size,"scB_count_agegrp_2001.txt",sep=""),sep="\t")
count_agegrp_2010=read.table(file=paste(n_size,"scB_count_agegrp_2010.txt",sep=""),sep="\t")

# Model convergency
model_c=matrix(0,nrow=n_sim,ncol=16)

# Save chosen model for each sample
model_chosen=NULL

# Predict survival at 5 years
timept=5

# significance level
alpha=0.05

for (k in 1:n_sim) {
setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo1/UM_SeminarioFev2017/Samples/2000/B")
model=NULL
model_1=model_2=model_3=model_4=model_5=model_6=NULL
model_7=model_8=model_9=model_10=model_11=model_12=NULL
model_13=model_14=model_15=model_16=NULL

print("Simulation n:")
print(k)
sample=read.table(file=paste(n_size,"scB_sample_",k,".txt",sep=""),sep="\t")
sample$age_yydx=sample$age_c*sample$yydx_c

sample_2001=sample[sample$yydx==2001,]
sample_2010=sample[sample$yydx==2010,]

# Flag if there is any age group with zero count
flag=0
for (j in 1:5) {
if (count_agegrp[k,j]==0) {flag=1}
}
flag2001=0
for (j in 1:5) {
if (count_agegrp_2001[k,j]==0) {flag2001=1}
}
flag2010=0
for (j in 1:5) {
if (count_agegrp_2010[k,j]==0) {flag2010=1}
}

# Model 1 vs Model 2 (test interaction age*yydx)
model_1=model1(sample)
model_2=model2(sample)

if (is.null(model_1$code)==F) {
if (model_1$code==1) {
model_c[k,1]=1
ll_1=model_1$loglik
```

```
}
}
df1=model_1$n.par
if(is.na(ll_1)) ll_1=-1e6

if (is.null(model_2$code)==F) {
if (model_2$code==1) {
model_c[k,2]=1
ll_2=model_2$loglik
}
}
df2=model_2$n.par
if(is.na(ll_2)) ll_2=-1e6

# Check if interaction age*yydx is significant
# (compare M1 with M2)
ll=2*(ll_1-ll_2)
p12=pchisq(ll, df=df1-df2,lower.tail = F)
print(p12)

if (p12<0.05) {
print("interaction is significant")
model_3=model3(sample)
model_4=model4(sample)
model_6=model6(sample)

if (is.null(model_3$code)==F) {
if (model_3$code==1) {
model_c[k,3]=1
ll_3=model_3$loglik
}
}
df3=model_3$n.par
if(is.na(ll_3)) ll_3=-1e6

if (is.null(model_4$code)==F) {
if (model_4$code==1) {
model_c[k,4]=1
ll_4=model_4$loglik
}
}
df4=model_4$n.par
if(is.na(ll_4)) ll_4=-1e6

if (is.null(model_6$code)==F) {
if (model_6$code==1) {
model_c[k,6]=1
ll_6=model_6$loglik
}
}
df6=model_6$n.par
if(is.na(ll_6)) ll_6=-1e6

# Compare models
# Test TD yydx
ll=2*(ll_2-ll_3)
p2_3=pchisq(ll, df=df2-df3,lower.tail = F)

# Test TD age
ll=2*(ll_2-ll_4)
p2_4=pchisq(ll, df=df2-df4,lower.tail = F)

# Test NL age
ll=2*(ll_2-ll_6)
p2_6=pchisq(ll, df=df2-df6,lower.tail = F)

max_p=max(p2_3,p2_4,p2_6)

if (max_p<0.05) {
model_chosen[k]="M1"
model=model_2
} else if (max_p>=alpha) {
model_5=model5(sample)
model_6=model6(sample)
model_7=model7(sample)
model_8=model8(sample)
model_9=model9(sample)

if (is.null(model_5$code)==F) {
if (model_5$code==1) {
model_c[k,5]=1
ll_5=model_5$loglik
}
}
df5=model_5$n.par
if(is.na(ll_5)) ll_5=-1e6
```

```
if (is.null(model_6$code)==F) {
if (model_6$code==1) {
model_c[k,6]=1
ll_6=model_6$loglik
}
}
df6=model_6$n.par
if(is.na(ll_6)) ll_6=-1e6

if (is.null(model_7$code)==F) {
if (model_7$code==1) {
model_c[k,7]=1
ll_7=model_7$loglik
}
}
df7=model_7$n.par
if(is.na(ll_7)) ll_7=-1e6

if (is.null(model_8$code)==F) {
if (model_8$code==1) {
model_c[k,8]=1
ll_8=model_8$loglik
}
}
df8=model_8$n.par
if(is.na(ll_8)) ll_8=-1e6

if (is.null(model_9$code)==F) {
if (model_9$code==1) {
model_c[k,9]=1
ll_9=model_9$loglik
}
}
df9=model_9$n.par
if(is.na(ll_9)) ll_9=-1e6

if(max_p==p2_3) {
# Compare M3/M5
ll=2*(ll_3-ll_5)
p3_5=pchisq(ll, df=df3-df5,lower.tail = F)
# Compare M3/M7
ll=2*(ll_3-ll_7)
p3_7=pchisq(ll, df=df3-df7,lower.tail = F)
max_p2=max(p3_5,p3_7)
if (max_p2<alpha) {
print("M3")
model=model_3
model_chosen[k]="M3"
} else {
if (max_p2==p3_5) {
# Compare M5/M9
ll=2*(ll_5-ll_9)
p5_9=pchisq(ll, df=df5-df9,lower.tail = F)
if (p5_9>=alpha) {
print("M9")
model=model_9
model_chosen[k]="M9"
} else {
print("M5")
model=model_5
model_chosen[k]="M5"
}
}
if (max_p2==p3_7) {
# Compare M7/M9
ll=2*(ll_7-ll_9)
p7_9=pchisq(ll, df=df7-df9,lower.tail = F)
if (p7_9>=alpha) {
print("M9")
model=model_9
model_chosen[k]="M9"
} else {
print("M7")
model=model_7
model_chosen[k]="M7"
}
}
}
}
if(max_p==p2_4) {
# Compare M4/M5
ll=2*(ll_4-ll_5)
p4_5=pchisq(ll, df=df4-df5,lower.tail = F)
# Compare M4/M8
ll=2*(ll_4-ll_8)
p4_8=pchisq(ll, df=df4-df8,lower.tail = F)
max_p2=max(p4_5,p4_8)
```

```
if (max_p2<alpha) {
print("M4")
model=model_4
model_chosen[k]="M4"
} else {
if (max_p2==p4_5) {
# Compare M5/M9
ll=2*(ll_5-ll_9)
p5_9=pchisq(ll, df=df5-df9,lower.tail = F)
if (p5_9>=alpha) {
print("M9")
model=model_9
model_chosen[k]="M9"
} else {
print("M5")
model=model_5
model_chosen[k]="M5"
}
}
if (max_p2==p4_8) {
# Compare M8/M9
ll=2*(ll_8-ll_9)
p8_9=pchisq(ll, df=df8-df9,lower.tail = F)
if (p8_9>=alpha) {
print("M9")
model=model_9
model_chosen[k]="M9"
} else {
print("M8")
model=model_8
model_chosen[k]="M8"
}
}
}
}
if(max_p==p2_6) {
# Compare M6/M7
ll=2*(ll_6-ll_7)
p6_7=pchisq(ll, df=df6-df7,lower.tail = F)
# Compare M6/M8
ll=2*(ll_6-ll_8)
p6_8=pchisq(ll, df=df6-df8,lower.tail = F)
max_p2=max(p6_7,p6_8)
if (max_p2<alpha) {
print("M6")
model=model_6
model_chosen[k]="M6"
} else {
if (max_p2==p6_7) {
# Compare M7/M9
ll=2*(ll_7-ll_9)
p7_9=pchisq(ll, df=df7-df9,lower.tail = F)
if (p7_9>=alpha) {
print("M9")
model=model_9
model_chosen[k]="M9"
} else {
print("M7")
model=model_7
model_chosen[k]="M7"
}
}
if (max_p2==p6_8) {
# Compare M8/M9
ll=2*(ll_8-ll_9)
p8_9=pchisq(ll, df=df8-df9,lower.tail = F)
if (p8_9>=alpha) {
print("M9")
model=model_9
model_chosen[k]="M9"
} else {
print("M8")
model=model_8
model_chosen[k]="M8"
}
}
}
}
}
} else if (p12>=alpha) {
print("interaction not significant")
model_10=model10(sample)
model_11=model11(sample)
model_13=model13(sample)

if (is.null(model_10$code)==F) {
if (model_10$code==1) {
```

```
model_c[k,10]=1
ll_10=model_10$loglik
}
}
df10=model_10$n.par
if(is.na(ll_10)) ll_10=-1e6

if (is.null(model_11$code)==F) {
if (model_11$code==1) {
model_c[k,11]=1
ll_11=model_11$loglik
}
}
df11=model_11$n.par
if(is.na(ll_11)) ll_11=-1e6

if (is.null(model_13$code)==F) {
if (model_13$code==1) {
model_c[k,13]=1
ll_13=model_13$loglik
}
}
df13=model_13$n.par
if(is.na(ll_13)) ll_13=-1e6

# Compare models
# Test TD yydx
ll=2*(ll_2-ll_10)
p2_10=pchisq(ll, df=df2-df10,lower.tail = F)

# Test TD age
ll=2*(ll_2-ll_11)
p2_11=pchisq(ll, df=df2-df11,lower.tail = F)

# Test NL age
ll=2*(ll_2-ll_13)
p2_13=pchisq(ll, df=df2-df13,lower.tail = F)

max_p=max(p2_10,p2_11,p2_13)

if (max_p<alpha) {
model_chosen[k]="M2"
model=model_2
} else if (max_p>=alpha) {
model_12=model12(sample)
model_13=model13(sample)
model_14=model14(sample)
model_15=model15(sample)
model_16=model16(sample)

if (is.null(model_12$code)==F) {
if (model_12$code==1) {
model_c[k,12]=1
ll_12=model_12$loglik
}
}
df12=model_12$n.par
if(is.na(ll_12)) ll_12=-1e6

if (is.null(model_13$code)==F) {
if (model_13$code==1) {
model_c[k,13]=1
ll_13=model_13$loglik
}
}
df13=model_13$n.par
if(is.na(ll_13)) ll_13=-1e6

if (is.null(model_14$code)==F) {
if (model_14$code==1) {
model_c[k,14]=1
ll_14=model_14$loglik
}
}
df14=model_14$n.par
if(is.na(ll_14)) ll_14=-1e6

if (is.null(model_15$code)==F) {
if (model_15$code==1) {
model_c[k,15]=1
ll_15=model_15$loglik
}
}
df15=model_15$n.par
if(is.na(ll_15)) ll_15=-1e6

if (is.null(model_16$code)==F) {
```

```
if (model_16$code==1) {
model_c[k,16]=1
ll_16=model_16$loglik
}
}
df16=model_16$n.par
if(is.na(ll_16)) ll_16=-1e6

if (max_p==p2_10) {
print(max_p)
# Compare M10/M12
ll=2*(ll_10-ll_12)
p10_12=pchisq(ll, df=df10-df12,lower.tail = F)
# Compare M10/M14
ll=2*(ll_10-ll_14)
p10_14=pchisq(ll, df=df10-df14,lower.tail = F)
max_p2=max(p10_12,p10_14)
if (max_p2<alpha) {
print("M10")
model=model_10
model_chosen[k]="M10"
} else {
if (max_p2==p10_12) {
# Compare M12/M16
ll=2*(ll_12-ll_16)
p12_16=pchisq(ll, df=df12-df16,lower.tail = F)
if (p12_16>=alpha) {
print("M16")
model=model_16
model_chosen[k]="M16"
} else {
print("M12")
model=model_12
model_chosen[k]="M12"
}
}
if (max_p2==p10_14) {
# Compare M14/M16
ll=2*(ll_14-ll_16)
p14_16=pchisq(ll, df=df14-df16,lower.tail = F)
if (p14_16>=alpha) {
print("M16")
model=model_16
model_chosen[k]="M16"
} else {
print("M14")
model=model_14
model_chosen[k]="M14"
}
}
}
}
if(max_p==p2_11) {
# Compare M11/M12
ll=2*(ll_11-ll_12)
p11_12=pchisq(ll, df=df11-df12,lower.tail = F)
# Compare M11/M15
ll=2*(ll_11-ll_15)
p11_15=pchisq(ll, df=df11-df15,lower.tail = F)
max_p2=max(p11_12,p11_15)
if (max_p2<alpha) {
print("M11")
model=model_11
model_chosen[k]="M11"
} else {
if (max_p2==p11_12) {
# Compare M12/M16
ll=2*(ll_12-ll_16)
p12_16=pchisq(ll, df=df12-df16,lower.tail = F)
if (p12_16>=alpha) {
print("M16")
model=model_16
model_chosen[k]="M16"
} else {
print("M12")
model=model_12
model_chosen[k]="M12"
}
}
if (max_p2==p11_15) {
# Compare M15/M16
ll=2*(ll_15-ll_16)
p15_16=pchisq(ll, df=df15-df16,lower.tail = F)
if (p15_16>=alpha) {
print("M16")
model=model_16
model_chosen[k]="M16"
} else {
print("M15")
model=model_15
model_chosen[k]="M15"
}
}
}
}
if(max_p==p2_13) {
# Compare M13/M14
ll=2*(ll_13-ll_14)
p13_14=pchisq(ll, df=df13-df14,lower.tail = F)
# Compare M13/M15
ll=2*(ll_13-ll_15)
p13_15=pchisq(ll, df=df13-df15,lower.tail = F)
max_p2=max(p13_14,p13_15)
if (max_p2<alpha) {
print("M13")
model=model_13
model_chosen[k]="M13"
} else {
if (max_p2==p13_14) {
# Compare M14/M16
ll=2*(ll_14-ll_16)
p14_16=pchisq(ll, df=df14-df16,lower.tail = F)
if (p14_16>=alpha) {
print("M16")
model=model_16
model_chosen[k]="M16"
} else {
print("M14")
model=model_14
model_chosen[k]="M14"
}
}
if (max_p2==p13_15) {
# Compare M15/M16
ll=2*(ll_15-ll_16)
p15_16=pchisq(ll, df=df15-df16,lower.tail = F)
if (p15_16>=alpha) {
print("M16")
model=model_16
model_chosen[k]="M16"
} else {
print("M15")
model=model_15
model_chosen[k]="M15"
}
}
}
}
}
}


if (model_chosen[k]!="") {
print("Converged")
# Predict age-standardised survival MB1 2001-2010
if (flag==0) {
sample$Weight[sample$age_grp==1]=weights[1]/nrow(sample[sample$age_grp==1,])
sample$Weight[sample$age_grp==2]=weights[2]/nrow(sample[sample$age_grp==2,])
sample$Weight[sample$age_grp==3]=weights[3]/nrow(sample[sample$age_grp==3,])
sample$Weight[sample$age_grp==4]=weights[4]/nrow(sample[sample$age_grp==4,])
sample$Weight[sample$age_grp==5]=weights[5]/nrow(sample[sample$age_grp==5,])
} else {
sample$Weight=1/nrow(sample)
}
model_res_MB1=PredSurvPop(mydata=sample,mytime=timept,mymodel=model,colweight="Weight")
asns_MB1[k]=model_res_MB1$SNW
se_MB1[k]=model_res_MB1$SNW*model_res_MB1$Stderr.logS

# Predict age-standardised survival MB1 2001
if (flag2001==0) {
sample_2001$Weight[sample_2001$age_grp==1]=weights[1]/nrow(sample_2001[sample_2001$age_grp==1,])
sample_2001$Weight[sample_2001$age_grp==2]=weights[2]/nrow(sample_2001[sample_2001$age_grp==2,])
sample_2001$Weight[sample_2001$age_grp==3]=weights[3]/nrow(sample_2001[sample_2001$age_grp==3,])
sample_2001$Weight[sample_2001$age_grp==4]=weights[4]/nrow(sample_2001[sample_2001$age_grp==4,])
sample_2001$Weight[sample_2001$age_grp==5]=weights[5]/nrow(sample_2001[sample_2001$age_grp==5,])
} else {
sample_2001$Weight=1/nrow(sample_2001)
}
model_res_MB1_2001=PredSurvPop(mydata=sample_2001,mytime=timept,mymodel=model,colweight="Weight")
asns_MB1_2001[k]=model_res_MB1_2001$SNW
se_MB1_2001[k]=model_res_MB1_2001$SNW*model_res_MB1_2001$Stderr.logS

# Predict age-standardised survival MB1 2010
if (flag2010==0) {
```

```
sample_2010$Weight[sample_2010$age_grp==1]=weights[1]/nrow(sample_2010[sample_2010$age_grp==1,])
sample_2010$Weight[sample_2010$age_grp==2]=weights[2]/nrow(sample_2010[sample_2010$age_grp==2,])
sample_2010$Weight[sample_2010$age_grp==3]=weights[3]/nrow(sample_2010[sample_2010$age_grp==3,])
sample_2010$Weight[sample_2010$age_grp==4]=weights[4]/nrow(sample_2010[sample_2010$age_grp==4,])
sample_2010$Weight[sample_2010$age_grp==5]=weights[5]/nrow(sample_2010[sample_2010$age_grp==5,])
} else {
sample_2010$Weight=1/nrow(sample_2010)
}
model_res_MB1_2010=PredSurvPop(mydata=sample_2010,mytime=timept,mymodel=model,colweight="Weight")
asns_MB1_2010[k]=model_res_MB1_2010$SNW
se_MB1_2010[k]=model_res_MB1_2010$SNW*model_res_MB1_2010$Stderr.logS

# Predict survival by age
if (model_chosen[k]=="M1"|model_chosen[k]=="M3"|model_chosen[k]=="M4"|model_chosen[k]=="M5"|
    model_chosen[k]=="M6"|model_chosen[k]=="M7"|model_chosen[k]=="M8"|model_chosen[k]=="M9") {
s_temp <- predict.mexhaz(model, time.pts=timept, data.val = data.frame(age_c=agecr_vector,
age_c2=agecr_vector2,age_c3=agecr_vector3,age_knot=agecr_vectortr,yydx_c=0,
age_yydx=0),conf.int="none")
} else {
s_temp <- predict.mexhaz(model, time.pts=timept, data.val = data.frame(age_c=agecr_vector,
age_c2=agecr_vector2,age_c3=agecr_vector3,age_knot=agecr_vectortr,yydx_c=0),conf.int="none")
}
surv_by_age[k,] <- as.numeric(s_temp$results["surv"][,1])

# MB2
prop_years=NULL
for (m in 1:10) {
prop_years[m]=length(sample$yydx[sample$yydx==(2000+m)])/length(sample$yydx)
}
m=0
for (j in 1:10) {
for (i in 1:5) {
m=m+1
mydata_mb2[m,1]=m1_agec[i]
mydata_mb2[m,2]=m1_agec2[i]
mydata_mb2[m,3]=m1_agec3[i]
mydata_mb2[m,4]=m1_ageknot[i]
mydata_mb2[m,5]=weights[i]*prop_years[j]
mydata_mb2[m,6]=2000+j-2005
mydata_mb2[m,7]=(2000+j-2005)*m1_agec[i]
}
}
colnames(mydata_mb2)=c("age_c","age_c2","age_c3","age_knot","peso","yydx_c","age_yydx")

s_temp01=PredSurvPop(mydata=mydata_mb2_2001,mytime=timept,mymodel=model,colweight="peso")
s_temp02=PredSurvPop(mydata=mydata_mb2_2002,mytime=timept,mymodel=model,colweight="peso")
s_temp03=PredSurvPop(mydata=mydata_mb2_2003,mytime=timept,mymodel=model,colweight="peso")
s_temp04=PredSurvPop(mydata=mydata_mb2_2004,mytime=timept,mymodel=model,colweight="peso")
s_temp05=PredSurvPop(mydata=mydata_mb2_2005,mytime=timept,mymodel=model,colweight="peso")
s_temp06=PredSurvPop(mydata=mydata_mb2_2006,mytime=timept,mymodel=model,colweight="peso")
s_temp07=PredSurvPop(mydata=mydata_mb2_2007,mytime=timept,mymodel=model,colweight="peso")
s_temp08=PredSurvPop(mydata=mydata_mb2_2008,mytime=timept,mymodel=model,colweight="peso")
s_temp09=PredSurvPop(mydata=mydata_mb2_2009,mytime=timept,mymodel=model,colweight="peso")
s_temp10=PredSurvPop(mydata=mydata_mb2_2010,mytime=timept,mymodel=model,colweight="peso")

  asns_MB2[k,1]=s_temp01$SNW
  asns_MB2[k,2]=s_temp02$SNW
  asns_MB2[k,3]=s_temp03$SNW
  asns_MB2[k,4]=s_temp04$SNW
  asns_MB2[k,5]=s_temp05$SNW
  asns_MB2[k,6]=s_temp06$SNW
  asns_MB2[k,7]=s_temp07$SNW
  asns_MB2[k,8]=s_temp08$SNW
  asns_MB2[k,9]=s_temp09$SNW
  asns_MB2[k,10]=s_temp10$SNW

  se_MB2[k,1]=s_temp01$SNW*s_temp01$Stderr.logS
se_MB2[k,2]=s_temp02$SNW*s_temp02$Stderr.logS
se_MB2[k,3]=s_temp03$SNW*s_temp03$Stderr.logS
se_MB2[k,4]=s_temp04$SNW*s_temp04$Stderr.logS
se_MB2[k,5]=s_temp05$SNW*s_temp05$Stderr.logS
se_MB2[k,6]=s_temp06$SNW*s_temp06$Stderr.logS
se_MB2[k,7]=s_temp07$SNW*s_temp07$Stderr.logS
se_MB2[k,8]=s_temp08$SNW*s_temp08$Stderr.logS
se_MB2[k,9]=s_temp09$SNW*s_temp09$Stderr.logS
se_MB2[k,10]=s_temp10$SNW*s_temp10$Stderr.logS

s_temp=PredSurvPop(mydata=mydata_mb2,mytime=timept,mymodel=model,colweight="peso")
asns_MB2_ave[k]=s_temp$SNW
se_MB2_ave[k]=s_temp$SNW*s_temp$Stderr.logS
}

setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo1/Paper_simulations/Scenario B2")

write.table(model_c,file=paste(n_size,"scB2_model_c.txt"),sep="\t")
write.table(model_chosen,file=paste(n_size,"scB2_model_chosen.txt"),sep="\t")
write.table(surv_by_age,file=paste(n_size,"scB2_surv_by_age.txt"),sep="\t")
write.table(asns_MB1,file=paste(n_size,"scB2_ASNS_MB1.txt"),sep="\t")
write.table(asns_MB1_2001,file=paste(n_size,"scB2_ASNS_MB1_2001.txt"),sep="\t")
write.table(asns_MB1_2010,file=paste(n_size,"scB2_ASNS_MB1_2010.txt"),sep="\t")

write.table(se_MB1,file=paste(n_size,"scB2_stand_ASNS_MB1.txt"),sep="\t")
write.table(se_MB1_2001,file=paste(n_size,"scB2_stand_ASNS_MB1_2001.txt"),sep="\t")
write.table(se_MB1_2010,file=paste(n_size,"scB2_stand_ASNS_MB1_2010.txt"),sep="\t")

write.table(asns_MB2,file=paste(n_size,"scB2_ASNS_MB2.txt"),sep="\t")
write.table(se_MB2,file=paste(n_size,"scB2_stand_ASNS_MB2.txt"),sep="\t")

write.table(asns_MB2_ave,file=paste(n_size,"scB2_ASNS_MB2_ave.txt"),sep="\t")
write.table(se_MB2_ave,file=paste(n_size,"scB2_stand_ASNS_MB2_ave.txt"),sep="\t")
}

# FUNCTION Model 1 - NL age + TD age + yydx + yydx*age + TD yydx
model1 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
     nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
     nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
      nph(age_c+yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 1
}

# FUNCTION Model 2 - NL age + TD age + yydx + TD yydx
model2 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
     nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
     nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
      nph(age_c+yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 2
}

# FUNCTION Model 3 - NL age + TD age + yydx + yydx*age + PH yydx
```

```
model3 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
     nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
     nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
      nph(age_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 3
}

# FUNCTION Model 4 - NL age + PH age + yydx + yydx*age + TD yydx
model4 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
     nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
     nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx+
      nph(yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 4
}

# FUNCTION Model 5 - NL age + PH age + yydx + yydx*age + PH yydx
model5 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+age_yydx,
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 5
}


# FUNCTION Model 6 - LL age + TD age + yydx + yydx*age + TD yydx
model6 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(age_c+yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 6
}

# FUNCTION Model 7 - LL age + TD age + yydx + yydx*age + PH yydx
model7 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(age_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 7
}

# FUNCTION Model 8 - LL age + TD age + yydx + yydx*age + PH yydx
model8 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
```

```
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx+nph(yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 8
}

# FUNCTION Model 9 - LL age + PH age + yydx + yydx*age + PH yydx
model9 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+age_yydx,
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 9
}

# FUNCTION Model 10 - NL age + TD age + yydx + PH yydx
model10 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
     nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
     nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
      nph(age_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
```

```
return(model)
# end of model 10
}

# FUNCTION Model 11 - NL age + PH age + yydx + TD yydx
model11 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
     nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
     nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c+
      nph(yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 11
}

# FUNCTION Model 12 - NL age + PH age + yydx + PH yydx
model12 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+age_c2+age_c3+age_knot+yydx_c,
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 12
}

# FUNCTION Model 13 - LL age + TD age + yydx + yydx*age + TD yydx
model13 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(age_c+yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(age_c+yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 13
}

# FUNCTION Model 14 - LL age + TD age + yydx + PH yydx
model14 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(age_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(age_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 14
}

# FUNCTION Model 15 - LL age + TD age + yydx + PH yydx
model15 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(yydx_c),
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c+nph(yydx_c),
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 15
}

# FUNCTION Model 16 - LL age + PH age + yydx + PH yydx
model16 <- function(sample) {
conv=0
# Get initial parameters estimates using observed survival
tryCatch({
model.crude <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5),fnoptim = c("optim"))},
     error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
my.init <- round(model.crude$coef, 2)
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c,
     data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",
     verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
if (is.null(model$code)==F) {
if (model$code==1) {
conv=1
} else {
conv=0
}
} else {
conv=0
}
if (conv==0) {
tryCatch({
model <- mexhaz(formula=Surv(time=stime, event=event)~age_c+yydx_c,
      data=sample, base="exp.bs", degree=3, knots=c(.25,1.5), expected="popmort",fnoptim = c("optim"),
      verbose=0, n.gleg=50,init=my.init)}, error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
return(model)
# end of model 16
}

#########################################################################
library(relsurv)

setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo1/Paper_simulations")

set.seed(123456)

# Sample size
n_size=2000

# Number of simulations
n_sim=1000

# Survival matrices by age group
surv_agegrp_PP=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_PP_2001=matrix(NA,nrow=n_sim,ncol=5)
surv_agegrp_PP_2010=matrix(NA,nrow=n_sim,ncol=5)

# Standard error matrices by age group
stand_agegrp_PP=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_PP_2001=matrix(NA,nrow=n_sim,ncol=5)
stand_agegrp_PP_2010=matrix(NA,nrow=n_sim,ncol=5)

# Unstandardised survival vectors
surv_unstd_PP=surv_unstd_PP_2001=surv_unstd_PP_2010=NULL

# Unstandardised SE vectors
stand_unstd_PP=stand_unstd_PP_2001=stand_unstd_PP_2010=NULL

# Life table definitions for Pohar-Perme estimator (package relsurv)
rt_men=as.matrix(read.table(file="rt_men.txt", header=F, sep="\t"))
rt_women=as.matrix(read.table(file="rt_women.txt", header=F, sep="\t"))
lifetable=transrate(rt_men,rt_women,yearlim=c(1995,2015),int.length=1)

setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo1/Paper_simulations/Samples/B/2000")

timept=5*365.25

for (k in 1:n_sim) {
print("Simulation n:")
print(k)
sample=read.table(file=paste(n_size,"scB_sample_",k,".txt",sep=""),sep="\t")
sample_2001=sample[sample$yydx==2001,]
sample_2010=sample[sample$yydx==2010,]

###########################################################################
# Estimate survival and std error - PP

for (l in 1:5) {
selection=which(sample$age_grp==l)
if (length(selection)>0) {
ns=rs.surv(Surv(stime*365.25,event)~1+ratetable(age=age*365.25,sex=sex,year=year_sim),
           ratetable=lifetable,data=sample[selection,])
tryCatch({
surv_agegrp_PP[k,l]=summary(ns, times=timept)$surv},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
tryCatch({
stand_agegrp_PP[k,l]=summary(ns, times=timept)$std.err},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
selection=which(sample_2001$age_grp==l)
if (length(selection)>0) {
```

```
ns2001=rs.surv(Surv(stime*365.25,event)~1+ratetable(age=age*365.25,sex=sex,year=year_sim),
           ratetable=lifetable,data=sample_2001[selection,])
tryCatch({
surv_agegrp_PP_2001[k,l]=summary(ns2001, times=timept)$surv},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
tryCatch({
stand_agegrp_PP_2001[k,l]=summary(ns2001, times=timept)$std.err},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
selection=which(sample_2010$age_grp==l)
if (length(selection)>0) {
ns2010=rs.surv(Surv(stime*365.25,event)~1+ratetable(age=age*365.25,sex=sex,year=year_sim),
           ratetable=lifetable,data=sample_2010[selection,])
tryCatch({
surv_agegrp_PP_2010[k,l]=summary(ns2010, times=timept)$surv},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
tryCatch({
stand_agegrp_PP_2010[k,l]=summary(ns2010, times=timept)$std.err},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}
}

# Estimate unstandardised net survival and stderr estimates for the full samples
# 2001-2010
ns=rs.surv(Surv(stime*365.25,event)~1+ratetable(age=age*365.25,sex=sex,year=year_sim),
          ratetable=lifetable,data=sample)
tryCatch({
surv_unstd_PP[k]=summary(ns, times=timept)$surv},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
tryCatch({
stand_unstd_PP[k]=summary(ns, times=timept)$std.err},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})

# 2001
ns_2001=rs.surv(Surv(stime*365.25,event)~1+ratetable(age=age*365.25,sex=sex,year=year_sim),
          ratetable=lifetable,data=sample_2001)
tryCatch({
surv_unstd_PP_2001[k]=summary(ns_2001, times=timept)$surv},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
tryCatch({
stand_unstd_PP_2001[k]=summary(ns_2001, times=timept)$std.err},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})

# 2010
ns_2010=rs.surv(Surv(stime*365.25,event)~1+ratetable(age=age*365.25,sex=sex,year=year_sim),
          ratetable=lifetable,data=sample_2010)
tryCatch({
surv_unstd_PP_2010[k]=summary(ns_2010, times=timept)$surv},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
tryCatch({
stand_unstd_PP_2010[k]=summary(ns_2010, times=timept)$std.err},
   error=function(e){cat("ERROR :",conditionMessage(e), "\n")})
}

setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo1/Paper_simulations/Scenario B - PP")

write.table(surv_agegrp_PP,file=paste(n_size,"scB_surv_agegrp_PP.txt"),sep="\t")
write.table(surv_agegrp_PP_2001,file=paste(n_size,"scB_surv_agegrp_PP_2001.txt"),sep="\t")
write.table(surv_agegrp_PP_2010,file=paste(n_size,"scB_surv_agegrp_PP_2010.txt"),sep="\t")

write.table(stand_agegrp_PP,file=paste(n_size,"scB_stand_agegrp_PP.txt"),sep="\t")
write.table(stand_agegrp_PP_2001,file=paste(n_size,"scB_stand_agegrp_PP_2001.txt"),sep="\t")
write.table(stand_agegrp_PP_2010,file=paste(n_size,"scB_stand_agegrp_PP_2010.txt"),sep="\t")

write.table(surv_unstd_PP,file=paste(n_size,"scB_surv_unstd_PP.txt"),sep="\t")
write.table(surv_unstd_PP_2001,file=paste(n_size,"scB_surv_unstd_PP_2001.txt"),sep="\t")
write.table(surv_unstd_PP_2010,file=paste(n_size,"scB_surv_unstd_PP_2010.txt"),sep="\t")

write.table(stand_unstd_PP,file=paste(n_size,"scB_stand_unstd_PP.txt"),sep="\t")
write.table(stand_unstd_PP_2001,file=paste(n_size,"scB_stand_unstd_PP_2001.txt"),sep="\t")
write.table(stand_unstd_PP_2010,file=paste(n_size,"scB_stand_unstd_PP_2010.txt"),sep="\t")
```

```
#
# Simulation of excess hazard model based on the
# Simulation performed in the article by Bartlett 2015
#
# Extension of SMC-FCS for excess hazard models
#
# Survival times: h_E(t|X)=0.002exp(beta1.X1 + beta2.X2)
# (beta1, beta2) = (1, 1)
# Censoring times: exponential distribution with hazard h_C(t)=0.002
# Background mortality: exponential distribution with hazard h_P(t)=0.001
#
# X1 ~ Be(n,p=0.5)
#
# X2 ~ N(X1, 1)
#
# Missing X1 - MAR: p_miss_x1=1/(1+exp(-0.1*x2+0.11))
# Missing X2 - MCAR: p_miss_x2=0.3
#
# Number of simulations: 1000
#
library(survival)
library(mice)
library(mitools)
library(smcfcs)
library(mexhaz)
library(MASS)

source("C:\\Users\\ljant\\Documents\\Doutoramento - PDMA\\Artigo_MI\\SummaryFunc.R")
source("C:\\Users\\ljant\\Documents\\Doutoramento - PDMA\\Artigo_MI\\function_smcfcs_exchaz.R")

set.seed(123)
n=1000
beta1=1
beta2=1
n_sim=1000

estim_x1=estim_x2=NULL
estim_x1_cc=estim_x2_cc=NULL
estim_x1_fcs=estim_x2_fcs=NULL
estim_x1_smcfcs=estim_x2_smcfcs=NULL

cov_cc_beta1=cov_cc_beta2=NULL
cov_fcs_beta1=cov_fcs_beta2=NULL
cov_smcfcs_beta1=cov_smcfcs_beta2=NULL

prob_x1=0.5
h0_x2=0.002
h_cens=0.002
h_pop=0.001
sd_x2=1

# Number of imputations
m_imps=10

for (k in 1:n_sim) {
# Generate X1
x1 = rbinom(n, 1, prob_x1)

# Generate X2
x2=NULL
for (i in 1:n) {
x2[i] = rnorm(1, mean=x1[i], sd=sd_x2)
}

# Generate time to death cancer
u1=runif(n,0,1)
t_E=log(1-u1)/(-h0_x2*exp(beta1*x1+beta2*x2))

# Generate censoring times
u2=runif(n,0,1)
c=log(1-u2)/(-h_cens)

# Generate time to death other causes
u3=runif(n,0,1)
t_P=log(1-u3)/(-h_pop)

# Define time to event as minimum between time to death from cancer and time to death from other causes
st_event=pmin(t_E,t_P)

# Define survival time as minimum between censoring and time to event
st=pmin(c,st_event)

# Create censoring indicator
d=ifelse(c<st_event,0,1)

# Create fully observed dataset
data=cbind(x1,x2,st,d)
```

```
data=as.data.frame(data)

# Bind population mortality
data$rate=rep(h_pop,n)

# Calcultate the Nelson-Aalen cumulative hazard estimate
data$nach=nelsonaalen(data, st, d)

# Create missing values
index=seq(1,n,1)
p_miss_x1=1/(1+exp(+0.01*st-0.3))
#sum(p_miss_x1>0.5)
#boxplot(p_miss_x1)
#plot(st,p_miss_x1)
mx2=sample(index,size=n*0.3)
data_inc=data
data_inc[p_miss_x1>0.5,"x1"]=NA
data_inc[mx2,"x2"]=NA

# Full dataset analysis
data$x1=as.factor(data$x1)
fit=mexhaz(formula=Surv(time=st, event=d)~x1+x2,data=data,base="pw.cst",expected="rate",
 verbose=0, n.gleg=25, fnoptim="optim")

estim_x1[k]=fit$coefficients["x11"]
estim_x2[k]=fit$coefficients["x2"]

# Complete case analysis
data_inc=as.data.frame(data_inc)
# Copy to other matrix since in smcfcs x1 must not be a factor
data_fcs=data_inc
data_fcs$x1=as.factor(data_inc$x1)

fit_cc=mexhaz(formula=Surv(time=st, event=d)~x1+x2,data=data_fcs,base="pw.cst",expected="rate",
  verbose=0, n.gleg=25, fnoptim="optim")

estim_x1_cc[k]=fit_cc$coefficients["x11"]
estim_x2_cc[k]=fit_cc$coefficients["x2"]

# Estimate coverage
lw1=up1=lw2=up2=NULL
# CI95% limits for beta1
lw1=confint(fit_cc)["x11",1]
up1=confint(fit_cc)["x11",2]

# CI95% limits for beta2
lw2=confint(fit_cc)["x2",1]
up2=confint(fit_cc)["x2",2]

cov_cc_beta1[k]=ifelse((lw1<beta1)&(beta1<up1),1,0)
cov_cc_beta2[k]=ifelse((lw2<beta2)&(beta2<up2),1,0)

# Impute missing values using FCS
# Select linear regression for X2 and logistic for X1
ini <- mice(data_inc, maxit = 0)
meth <- ini$meth
pred <- ini$pred
meth["x1"] <- "logreg"
meth["x2"] <- "norm"
pred["x1","st"]=0
pred["x2","st"]=0
imp=mice(data_fcs,meth=meth,pred=pred,m=m_imps)

analysis_imput_fcs <- with_mexhaz(imp)
combined_coef_fcs <- combine_mexhaz(analysis_imput_fcs)

estim_x1_fcs[k]=combined_coef_fcs["x1","mean"]
estim_x2_fcs[k]=combined_coef_fcs["x2","mean"]

# Estimate coverage
lw1=up1=lw2=up2=NULL
# CI95% limits for beta1
lw1=combined_coef_fcs["x1","lower"]
up1=combined_coef_fcs["x1","upper"]

# CI95% limits for beta2
lw2=combined_coef_fcs["x2","lower"]
up2=combined_coef_fcs["x2","upper"]

cov_fcs_beta1[k]=ifelse((lw1<beta1)&(beta1<up1),1,0)
cov_fcs_beta2[k]=ifelse((lw2<beta2)&(beta2<up2),1,0)

# Impute missing values using SMC-FCS
imp_smcfcs=smcfcs_exchaz(data_inc,m_imps)
analysis_imput_smcfcs=with_mexhaz_smcfcs(imp_smcfcs,m_imps)
combined_coef_smcfcs <- combine_mexhaz(analysis_imput_smcfcs)
```

```
estim_x1_smcfcs[k]=combined_coef_smcfcs["x1","mean"]
estim_x2_smcfcs[k]=combined_coef_smcfcs["x2","mean"]

# Estimate coverage
lw1=up1=lw2=up2=NULL
# CI95% limits for beta1
lw1=combined_coef_smcfcs["x1","lower"]
up1=combined_coef_smcfcs["x1","upper"]

# CI95% limits for beta2
lw2=combined_coef_smcfcs["x2","lower"]
up2=combined_coef_smcfcs["x2","upper"]

cov_smcfcs_beta1[k]=ifelse((lw1<beta1)&(beta1<up1),1,0)
cov_smcfcs_beta2[k]=ifelse((lw2<beta2)&(beta2<up2),1,0)
}

boxplot(estim_x1_cc,estim_x1_fcs,estim_x1_smcfcs,
  estim_x2_cc,estim_x2_fcs,estim_x2_smcfcs,axt="n")
axis(1,at=c(1,2,3,4,5,6),c("CC-X1","FCS-X1","SMC-FCS-X1","CC-X2","FCS-X2","SMC-FCS-X2"))
abline(h=1,col=2)

# CC
emp_cov_cc_b1=round(sum(na.omit(cov_cc_beta1))/sum(is.na(cov_cc_beta1)==F)*100,1)
emp_cov_cc_b2=round(sum(na.omit(cov_cc_beta2))/sum(is.na(cov_cc_beta2)==F)*100,1)

# FCS
emp_cov_fcs_b1=round(sum(na.omit(cov_fcs_beta1))/sum(is.na(cov_fcs_beta1)==F)*100,1)
emp_cov_fcs_b2=round(sum(na.omit(cov_fcs_beta2))/sum(is.na(cov_fcs_beta2)==F)*100,1)

# SMC-FCS
emp_cov_smcfcs_b1=round(sum(na.omit(cov_smcfcs_beta1))/sum(is.na(cov_smcfcs_beta1)==F)*100,1)
emp_cov_smcfcs_b2=round(sum(na.omit(cov_smcfcs_beta2))/sum(is.na(cov_smcfcs_beta2)==F)*100,1)

smcfcs_exchaz <- function(data,m) {
# m - number of imputations
# n - number of rows of matrix
n=dim(data)[1]

# Completed datasets
data_compl=list()

max_iter_reject=1000
n_iter=10

# Population hazard must be stored in variable "rate"
data$lambda_p=data$rate

# Cumulative population hazard
# (in this first example, population hazard is constant)
data$cum_lambda_p=data$rate*data$st

# store which rows in the matrix have x1 or x2 missing
row_miss_x1=which(is.na(data$x1)==T)
row_miss_x2=which(is.na(data$x2)==T)

# Fill in all missing values with starting value
# Mode for the categorical variable (x1)
# Mean for the continuos variable (x2)
mode_x1=getmode(na.omit(data$x1))
mean_x2=mean(na.omit(data$x2))

data1=data
data1$x1[is.na(data1$x1)==T]=mode_x1
data1$x2[is.na(data1$x2)==T]=mean_x2

# Count number of cases that have missing values for each variable
n_miss1=length(row_miss_x1)
n_miss2=length(row_miss_x2)

# Keep imputations from iterations
imputed_x1=matrix(NA,nrow=n_miss1,ncol=n_iter)
imputed_x2=matrix(NA,nrow=n_miss2,ncol=n_iter)

for (imp in 1:m) {
for (l in 1:n_iter) {
# Fit the substantive model of interest to the completed dataset
fit1=mexhaz(formula=Surv(time=st, event=d)~x1+x2,data=data1,base="pw.cst",
                expected="rate",verbose=0, n.gleg=25, fnoptim="optim")

# Store estimated coefficients and var-cov matrix from the substantive model
beta_means=c(fit1$coefficients["x1"],fit1$coefficients["x2"])
beta_vcov=fit1$vcov[2:3,2:3]

# Draw random values of the model parameters from a normal multivariate distribution
# with mean equal to parameters estimates and var-cov matrix estimated from the model
betas_subst=mvrnorm(n=1,mu=beta_means,Sigma=beta_vcov)
# The excess hazard baseline is considered piecewise constant (with only one time interval)
data1$lambda_baseline=exp(fit1$coefficients[1])
data1$cum_lambda_baseline=exp(fit1$coefficients[1])*data1$st

# Fit imputation model to x1 conditioned on x2
fit_impmod_x1=glm(x1~x2,family=binomial(link="logit"),data=data1)
fit_impmod_x1_coeff=fit_impmod_x1$coefficients
fit_impmod_x1_vcov=stats::vcov(fit_impmod_x1)

# Draw random values of the imputation model parameters from a normal multivariate
# distribution with mean equal to parameters estimates and var-cov matrix estimated
# from the fitted imputation model
betas_impmod_x1=mvrnorm(n=1,mu=fit_impmod_x1_coeff,Sigma=fit_impmod_x1_vcov)

# For X1
for (j in 1:n_miss1) {
# Draw random imputation for X1
# First, calculate probability from logistic regression model
k=row_miss_x1[j]
pr_x1=1/(1+exp(-(betas_impmod_x1[1]+betas_impmod_x1[2]*x2[k])))

count=0
reject=1
while(reject==1 & count<max_iter_reject) {
count=count+1
# Draw from binomial distribution
x1_imp=rbinom(1,1,pr_x1)

reject = check_comp(x1_imp,data1$x2[k],data1$lambda_p[k],data1$cum_lambda_p[k],
                data1$lambda_baseline[k],data1$cum_lambda_baseline[k],
             data1$d[k],betas_subst)
}
# Save iteration
imputed_x1[j,l]=x1_imp

# Replace x1 values in data by newly imputed ones
data1[k,"x1"]=x1_imp
}

# Fit imputation model to x2 conditioned on x1
fit_impmod_x2=lm(x2~x1,data=data1)
fit_impmod_x2_coeff=fit_impmod_x2$coefficients
fit_impmod_x2_vcov=stats::vcov(fit_impmod_x2)
fit_impmod_x2_resid_sd=(summary(fit_impmod_x2)$sigma)

# Draw random values of the imputation model parameters from a normal multivariate
# distribution with mean equal to parameters estimates and var-cov matrix estimated
# from the fitted imputation model
betas_impmod_x2=mvrnorm(n=1,mu=fit_impmod_x2_coeff,Sigma=fit_impmod_x2_vcov)

# For X2
for (j in 1:n_miss2) {
# Draw random imputation for X2 (from linear model)
k=row_miss_x2[j]
count=0
reject=1
while(reject==1 & count<max_iter_reject) {
count=count+1
x2_imp=as.numeric(betas_impmod_x2[1]+betas_impmod_x2[2]*data1$x1[k]+
rnorm(n=1, mean = 0, sd = fit_impmod_x2_resid_sd))

reject = check_comp(data1$x1[k],x2_imp,data1$lambda_p[k],data1$cum_lambda_p[k],
                    data1$lambda_baseline[k],data1$cum_lambda_baseline[k],
                data1$d[k],betas_subst)
}
# Save iteration
imputed_x2[j,l]=x2_imp

# Replace x1 values in data by newly imputed ones
data1[k,"x2"]=x2_imp
# print(paste("count=",count))
# print(paste("x1=",x1_imp))
# print(paste("d=",data1$d[k]))
}
}

# Save completed dataset
data_compl[[imp]]=data1
}
return(data_compl)
}


# Function to calculate mode of a vector
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Function for sample rejection - X1
check_comp <- function(x1,x2,lambda_p,cum_lambda_p,lambda_baseline,cum_lambda_baseline,
        d,betas_subst) {
if (d==0) {
#print("0")
c_lim = exp(-cum_lambda_p) * exp(-cum_lambda_baseline*
   exp(betas_subst[1]*x1+betas_subst[2]*x2))
u=runif(1)
reject=ifelse(u<=c_lim,0,1)
} else if (d==1) {
#print("1")
A=cum_lambda_baseline*lambda_p/lambda_baseline
B=exp(betas_subst[1]*x1+betas_subst[2]*x2)
c_lim=((lambda_p+lambda_baseline*B)*exp(-cum_lambda_p-cum_lambda_baseline*B))/
((lambda_baseline/cum_lambda_baseline)*exp(-cum_lambda_p-1+A))
u=runif(1)
reject=ifelse(u<=c_lim,0,1)
}
return(reject)
}

with_mexhaz_smcfcs <- function(imputed,m)  {
coef=matrix(NA,ncol=m,nrow=4)
rownames(coef)=c("x1","x2","x1_std","x2_std")
for (i in 1:m) {
data=imputed[[i]]
fit=mexhaz(formula=Surv(time=st, event=d)~x1+x2,data=data,
     base="pw.cst",expected="rate",
     verbose=0, n.gleg=25, fnoptim="optim")

coef["x1",i]=fit$coefficients["x1"]
coef["x2",i]=fit$coefficients["x2"]

coef["x1_std",i]=fit$std.errors["x1"]
coef["x2_std",i]=fit$std.errors["x2"]
}
return(coef)
}

combine_mexhaz <- function(coef) {
ncoef=(dim(coef)[1])/2
m=dim(coef)[2]
results=matrix(NA,ncol=4,nrow=ncoef)
rownames(results)=c("x1","x2")
colnames(results)=c("mean","std.err","lower","upper")
for (i in 1:ncoef) {
results[i,1]=mean(coef[i,])
U=mean(coef[(i+ncoef),]^2)
B=1/(m-1)*sum((coef[i,]-mean(coef[i,]))^2)
results[i,2]=sqrt(U+(1+1/m)*B)
# Calculate confidence interval limits
df=floor((m-1)*(1+U/B)^2)
t=qt(0.975,df)
results[i,3]=results[i,1]-t*results[i,2]
results[i,4]=results[i,1]+t*results[i,2]
}
return(results)
}
```

```
##########################################################################################
# Colorectal cancer in the North region of Portugal
#
# Article on Multiple Imputation extending SMC-FCS to accomodate excess hazard models
# Excess hazard modelling
#
# Complete case analysis, FCS, SMC-FCS
#
# Covariables: age, sex, EDI, extent
#

# Load necessary libraries
library(MASS)
library(relsurv)
library(mexhaz)
library(dplyr)
library(mice)
library(mitools)

set.seed(1234)

source("C:\\Users\\ljant\\Documents\\Doutoramento - PDMA\\Artigo_MI\\SummaryFunc.R")
source("C:\\Users\\ljant\\Documents\\Doutoramento - PDMA\\Artigo_MI\\functions_application.R")
source("C:\\Users\\ljant\\Documents\\Doutoramento - PDMA\\Artigo_MI\\functions_application_smcfcs_v4.R")

# Set working directory
setwd("C://Users/ljant/Documents/Doutoramento - PDMA/Artigo_MI")

# Load initial database
bd_original<-read.csv2("Base_CRC_2010-2012_actual.csv",header=TRUE,sep=";")

bd_original$survtime_yrs=bd_original$survtime/365.25

# Create variable age group
bd_original$grupo_et[(bd_original$Idade>=15)&(bd_original$Idade<=44)]=1
bd_original$grupo_et[(bd_original$Idade>=45)&(bd_original$Idade<=54)]=2
bd_original$grupo_et[(bd_original$Idade>=55)&(bd_original$Idade<=64)]=3
bd_original$grupo_et[(bd_original$Idade>=65)&(bd_original$Idade<=74)]=4
bd_original$grupo_et[(bd_original$Idade>=75)&(bd_original$Idade<=120)]=5

# Code sex
bd_original$sex[bd_original$Sexo=="Masculino"]=1
bd_original$sex[bd_original$Sexo=="Feminino"]=2

bd_original$dep[bd_original$EDI==1]="edi1"
bd_original$dep[bd_original$EDI==2]="edi2"
bd_original$dep[bd_original$EDI==3]="edi3"
bd_original$dep[bd_original$EDI==4]="edi4"
bd_original$dep[bd_original$EDI==5]="edi5"

# Read deprivation-specific life tables
rt_men1<-as.matrix(read.table(file="rt_men_edi1.txt", header=F, sep="\t"))
rt_women1<-as.matrix(read.table(file="rt_women_edi1.txt", header=F, sep="\t"))
rt_men2<-as.matrix(read.table(file="rt_men_edi2.txt", header=F, sep="\t"))
rt_women2<-as.matrix(read.table(file="rt_women_edi2.txt", header=F, sep="\t"))
rt_men3<-as.matrix(read.table(file="rt_men_edi3.txt", header=F, sep="\t"))
rt_women3<-as.matrix(read.table(file="rt_women_edi3.txt", header=F, sep="\t"))
rt_men4<-as.matrix(read.table(file="rt_men_edi4.txt", header=F, sep="\t"))
rt_women4<-as.matrix(read.table(file="rt_women_edi4.txt", header=F, sep="\t"))
rt_men5<-as.matrix(read.table(file="rt_men_edi5.txt", header=F, sep="\t"))
rt_women5<-as.matrix(read.table(file="rt_women_edi5.txt", header=F, sep="\t"))

# Deprivation-specific life tables
lifetable1<-transrate(rt_men1,rt_women1,yearlim=c(2001,2017),int.length=1)
lifetable2<-transrate(rt_men2,rt_women2,yearlim=c(2001,2017),int.length=1)
lifetable3<-transrate(rt_men1,rt_women3,yearlim=c(2001,2017),int.length=1)
lifetable4<-transrate(rt_men4,rt_women4,yearlim=c(2001,2017),int.length=1)
lifetable5<-transrate(rt_men5,rt_women5,yearlim=c(2001,2017),int.length=1)

lifetable_edi <- joinrate(list(edi1=lifetable1,edi2=lifetable2,edi3=lifetable3,edi4=lifetable4,
edi5=lifetable5),dim.name="deprivation")

# Transform probability of survival in mortality rate
rate_men1=-log(rt_men1)
rate_men2=-log(rt_men2)
rate_men3=-log(rt_men3)
rate_men4=-log(rt_men4)
rate_men5=-log(rt_men5)
rate_women1=-log(rt_women1)
rate_women2=-log(rt_women2)
rate_women3=-log(rt_women3)
rate_women4=-log(rt_women4)
rate_women5=-log(rt_women5)

# Attribute column names to mortality matrix
colnames(rate_men1)=seq(2001,2017,1)
colnames(rate_men2)=seq(2001,2017,1)
colnames(rate_men3)=seq(2001,2017,1)
colnames(rate_men4)=seq(2001,2017,1)
colnames(rate_men5)=seq(2001,2017,1)
colnames(rate_women1)=seq(2001,2017,1)
colnames(rate_women2)=seq(2001,2017,1)
colnames(rate_women3)=seq(2001,2017,1)
colnames(rate_women4)=seq(2001,2017,1)
colnames(rate_women5)=seq(2001,2017,1)

# Attribute row names to mortality matrix
rownames(rate_men1)=seq(0,99,1)
rownames(rate_men2)=seq(0,99,1)
rownames(rate_men3)=seq(0,99,1)
rownames(rate_men4)=seq(0,99,1)
rownames(rate_men5)=seq(0,99,1)
rownames(rate_women1)=seq(0,99,1)
rownames(rate_women2)=seq(0,99,1)
rownames(rate_women3)=seq(0,99,1)
rownames(rate_women4)=seq(0,99,1)
rownames(rate_women5)=seq(0,99,1)

# Apply exclusion criteria
# Limit to patients with:
#  - age at diagnosis < 95
#  - known status
#  - survival time > 0
#  - known EDI

bd_surv = bd_original %>% filter(Idade<95) %>% filter(status!=9) %>%
  filter(survtime>0) %>% filter(EDI!=9)

dim(bd_surv)

# Read and append population mortality rate
n=dim(bd_surv)[1]
for (i in 1:n) {
bd_surv$age_exit[i]=floor(bd_surv$survtime_yrs[i])+bd_surv$Idade[i]
bd_surv$year_exit[i]=floor(bd_surv$survtime_yrs[i])+bd_surv$anodiag[i]
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi1") {
bd_surv$popmort_spec[i]=
rate_men1[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi2") {
bd_surv$popmort_spec[i]=
rate_men2[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi3") {
bd_surv$popmort_spec[i]=
rate_men3[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi4") {
bd_surv$popmort_spec[i]=
rate_men4[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi5") {
bd_surv$popmort_spec[i]=
rate_men5[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}

if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi1") {
bd_surv$popmort_spec[i]=
rate_women1[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi2") {
bd_surv$popmort_spec[i]=
rate_women2[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi3") {
bd_surv$popmort_spec[i]=
rate_women3[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi4") {
bd_surv$popmort_spec[i]=
rate_women4[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi5") {
bd_surv$popmort_spec[i]=
rate_women5[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
}


# Calculate and append cumulative population hazard
# Cum pop hazard = SUM (pop haz (age_i) * time)
for (i in 1:n) {
bd_surv$yrs_lived[i]=bd_surv$age_exit[i]-bd_surv$Idade[i]+1
bd_surv$cum_popmort_spec[i]=0
```

```
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi1") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_men1[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi2") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_men2[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi3") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_men3[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi4") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_men4[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==1 & bd_surv$dep[i]=="edi5") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_men5[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}

if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi1") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_women1[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi2") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_women2[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi3") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_women3[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi4") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_women4[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
if(bd_surv$sex[i]==2 & bd_surv$dep[i]=="edi5") {
for (j in 1:bd_surv$yrs_lived[i]) {
bd_surv$cum_popmort_spec[i]=bd_surv$cum_popmort_spec[i]+
rate_women5[as.character(bd_surv$Idade[i]-1+j),as.character(bd_surv$anodiag[i]-1+j)]
}
}
}

# Distribution of EDI quintiles in the North region of Portugal
# (quintiles were obtained from the all country distribution)
weights_pop=c(0.09481,0.11785,0.21968,0.26554,0.30212)

# Define matrix for the weighted mortalities
rate_men_gen=matrix(NA,nrow=100,ncol=17)
rate_women_gen=matrix(NA,nrow=100,ncol=17)
colnames(rate_men_gen)=seq(2001,2017,1)
colnames(rate_women_gen)=seq(2001,2017,1)
rownames(rate_men_gen)=seq(0,99,1)
rownames(rate_women_gen)=seq(0,99,1)

nr=dim(rate_men_gen)[1]
nc=dim(rate_men_gen)[2]

# Calculate weighted mortalities to obtain general life tables
# (better to compare results from deprivation specific and general life tables)
for (i in 1:nr) {
for (j in 1:nc) {
rate_men_gen[i,j]=weights_pop[1]*rate_men1[i,j]+weights_pop[2]*rate_men2[i,j]+
 weights_pop[3]*rate_men3[i,j]+weights_pop[4]*rate_men4[i,j]+
 weights_pop[5]*rate_men5[i,j]
rate_women_gen[i,j]=weights_pop[1]*rate_women1[i,j]+weights_pop[2]*rate_women2[i,j]+
   weights_pop[3]*rate_women3[i,j]+weights_pop[4]*rate_women4[i,j]+
   weights_pop[5]*rate_women5[i,j]
}
}

n=dim(bd_surv)[1]
for (i in 1:n) {
if(bd_surv$sex[i]==1) {
bd_surv$popmort_gen_calc[i]=
rate_men_gen[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
if(bd_surv$sex[i]==2) {
bd_surv$popmort_gen_calc[i]=
rate_women_gen[as.character(bd_surv$age_exit[i]),as.character(bd_surv$year_exit[i])]
}
}

# Create dummy variables for EDI
bd_surv$edi2=bd_surv$edi3=bd_surv$edi4=bd_surv$edi5=0

bd_surv$edi2[bd_surv$EDI==2]=1
bd_surv$edi3[bd_surv$EDI==3]=1
bd_surv$edi4[bd_surv$EDI==4]=1
bd_surv$edi5[bd_surv$EDI==5]=1

# Create dummy variables for extent
bd_surv$extent2=bd_surv$extent3=bd_surv$extent4=0

bd_surv$extent2[bd_surv$extent==2]=1
bd_surv$extent3[bd_surv$extent==3]=1
bd_surv$extent4[bd_surv$extent==4]=1

# Create dummy variable for sex (women - reference category)
bd_surv$men=0
bd_surv$men[bd_surv$sex==1]=1

# Standardise age
c_age=70
s_age=12
bd_surv$age_z=(bd_surv$Idade-c_age)/s_age

####################################################################
#
# Estimation of net survival by EDI for the full cohort
# Compare survival curves using "log-rank type" test
#
####################################################################

pp_edi<-rs.surv(Surv(survtime,status==1) ~ EDI + ratetable(age=Idade*365.25,sex=sex,year=year,
deprivation=dep),ratetable=lifetable_edi,data=bd_surv,conf.type="log")

p=rs.diff(Surv(survtime,status==1) ~ EDI + ratetable(age=Idade*365.25,sex=sex,year=year,
deprivation=dep),ratetable=lifetable_edi,data=bd_surv)

plot(pp_edi,col=c(1,2,3,4,5),lty=c(1,2,3,4,5),xlim=c(0,5),
ylab="Net survival", xlab="Years since diagnosis",xscale=365.25,xaxs="i",yaxs="i",font.lab=2)
legend("bottomright",c("EDI1 (least deprived)","EDI2","EDI3","EDI4","EDI5 (most deprived)"),
lty=seq(1,5,1),col=seq(1,5,1),bty="n")
text(x=4*365,y=0.9,paste("p = ",round(as.numeric(p$p.value),4)),font=3)


####################################################################
#
# COMPLETE-CASE ANALYSIS
#
# Discard cases with missing extent
#
####################################################################

bd_surv_cc = bd_surv %>% filter(extent!=9)

dim(bd_surv_cc)

# Fit model with Linear age and PH effects of age, sex, edi and extent

model_cc <- mexhaz(formula=Surv(time=survtime_yrs, event=status)~age_z + men +
       edi2 + edi3 + edi4 + edi5 + extent2 + extent3 + extent4,
       data=bd_surv_cc, base="exp.bs", degree=3, knots=c(1), expected="popmort_spec",
       verbose=0, n.gleg=50)

summary2(model_cc)
model_cc

####################################################################
#
# MULTIPLE IMPUTATION USING FCS
#
# Imputation model: multinomial logistic regression model
# Covariates in imputation model: age, sex, tumour site,
# deprivation, basis of diagnosis,
#      cumulative hazard, event indicator
#
###################################################################

# Define number of imputations
n_imps=50

# Define number of iterations
n_iter=5

# Create dummy variable for tumour site
bd_surv$site[bd_surv$Top=="colon"]=0
bd_surv$site[bd_surv$Top=="rectum"]=1

# Create dummy variable for basis of diagnosis
bd_surv$mv=0
bd_surv$mv[bd_surv$basediag=="Histolgico"|
     bd_surv$basediag=="Citolgico"]=1

# Calcultate the Nelson-Aalen cumulative hazard estimate
bd_surv$nach=nelsonaalen(bd_surv, survtime, status)

bd_surv_fcs = bd_surv %>%
  ### SELECT A SUBSET OF VARIABLES TO WORK WITH
  select(
    age_z,
    men,
    site,
    EDI,
    mv,
    nach,
    status,
    survtime_yrs,
    extent,
    popmort_spec,
    cum_popmort_spec,
    year
  )

# Delete missing extent (9 --> NA)
bd_surv_fcs[bd_surv_fcs$extent==9,"extent"]=NA

# Define categorical variables as factors
bd_surv_fcs$EDI=as.factor(bd_surv_fcs$EDI)
bd_surv_fcs$extent=as.factor(bd_surv_fcs$extent)

# Initialise Multiple Imputation
ini <- mice(bd_surv_fcs, maxit = 0)

# Eliminate variables from the imputation model
pred <- ini$pred
pred[ ,"survtime_yrs"] <- 0
pred[ ,"popmort_spec"] <- 0
pred[ ,"cum_popmort_spec"] <- 0
pred[ ,"year"] <- 0

pred
imp_fcs <- mice(bd_surv_fcs, pred=pred, m=n_imps, maxit=n_iter)

coefs_fcs=with_mexhaz_fcs(imp_fcs)

comb_coef_fcs=combine_fcs(coefs_fcs)

comb_coef_fcs


##############################################################
#
# MULTIPLE IMPUTATION USING SMC-FCS
#
# Imputation model: multinomial logistic regression model
# Covariates in imputation model: age, sex, deprivation, mv, site
#
##############################################################

# Assumes that the above preparation code was run
bd_surv_smcfcs=bd_surv_fcs

n_iter=10
library(nnet)

n_iter=10
n_imps=50

imp_smcfcs = smcfcs_exchaz_app(bd_surv_smcfcs,n_imps,n_iter)
```

```
coefs_smcfcs=with_mexhaz_smcfcs_app(imp_smcfcs,m=n_imps)

comb_coef_smcfcs=combine_mexhaz_app(coefs_smcfcs)


smcfcs_exchaz_app <- function(data,m,n_iter) {
# m - number of imputations
# n - number of rows of matrix
n=dim(data)[1]

# Completed datasets
data_compl=list()

max_iter_reject=1000

# Population hazard must be stored in variable "popmort_spec"
data$lambda_p=data$popmort_spec

# Cumulative population hazard
# Read from data (calculated elsewhere)
data$cum_lambda_p=data$cum_popmort_spec

# store which rows in the matrix have extent missing
row_miss_ext=which(is.na(data$extent)==T)

# Count number of cases that have missing values for each variable
n_miss=length(row_miss_ext)

# Create dummy variables for EDI
data$edi2=data$edi3=data$edi4=data$edi5=0
data$edi2[data$EDI==2]=1
data$edi3[data$EDI==3]=1
data$edi4[data$EDI==4]=1
data$edi5[data$EDI==5]=1

# Fill in all missing values with starting value
# Predicted from a multinomial regression model with survtime as explanatory variable
# (should converge faster than using the mode)
fit_init=multinom(extent˜survtime_yrs,data=data[is.na(data$extent)==F,])

data1=data
for (j in 1:n_miss) {
k=row_miss_ext[j]
pred_ext=predict(fit_init,data1[k,])
data1$extent[k]=pred_ext
}

# Define extent as factor
data1$extent=as.factor(data1$extent)

# Keep imputations from iterations
imputed_ext=matrix(NA,nrow=n_miss,ncol=n_iter)

# Cycle for imputations
for (imp in 1:m) {
print(paste("imputation=",imp))
# Cycle for iteration within each imputation
for (l in 1:n_iter) {
print(paste("iter=",l))
# Fit the substantive model of interest to the completed dataset
fit1 = mexhaz(formula=Surv(time=survtime_yrs, event=status)˜age_z + men +
                edi2 + edi3 + edi4 + edi5 + extent +
  site + mv,
                    data=data1, base="exp.bs", degree=3, knots=c(1), expected="popmort_spec",
          verbose=0, n.gleg=50, fnoptim="optim")

# Store estimated coefficients for excess hazard baseline
gamma_means=c(fit1$coefficients["Intercept"],fit1$coefficients["BS3.1"],
  fit1$coefficients["BS3.2"],fit1$coefficients["BS3.3"],
  fit1$coefficients["BS3.4"])
gamma_vcov =fit1$vcov[1:5,1:5]

# Store estimated coefficients and var-cov matrix from the substantive model
beta_means=c(fit1$coefficients["age_z"],fit1$coefficients["men"],
 fit1$coefficients["edi2"],fit1$coefficients["edi3"],
 fit1$coefficients["edi4"],fit1$coefficients["edi5"],
 fit1$coefficients["extent2"],fit1$coefficients["extent3"],
 fit1$coefficients["extent4"],fit1$coefficients["site"],
 fit1$coefficients["mv"])
beta_vcov=fit1$vcov[6:16,6:16]

# Draw random values of the model parameters from a normal multivariate distribution
# with mean equal to parameters estimates and var-cov matrix estimated from the model
# Draw baseline parameters separately from covariates parameters
gamma_subst=mvrnorm(n=1,mu=gamma_means,Sigma=gamma_vcov)
betas_subst=mvrnorm(n=1,mu=beta_means,Sigma=beta_vcov)
# Use predict function from mexhaz to estimate excess hazard baseline
# and cumulative excess hazard baseline
fit_temp=fit1
coefs_temp=c(gamma_subst,betas_subst)
fit_temp$coefficients=coefs_temp

pred_baseline=predict(fit_temp, time.pts=data1$survtime_yrs,
 data.val = data.frame(age_z=0,men=0,edi2=0,edi3=0,
       edi4=0,edi5=0,extent=as.factor(1),site=0,mv=0),conf.int="none")

for (p in 1:n) {
data1$lambda_baseline[p]=pred_baseline$results$hazard[
pred_baseline$results$time.pts==data1$survtime_yrs[p]][1]
data1$cum_lambda_baseline[p]=-log(pred_baseline$results$surv[
pred_baseline$results$time.pts==data1$survtime_yrs[p]][1])
}

# Fit imputation model to extent conditioned on:
#  age_z, sex, deprivation, site, mv

# For Extent
y=data1$extent
ry=ifelse(is.na(bd_surv_smcfcs$extent)==F,T,F)

x = data1 %>%
   select(
age_z, men, site, edi2, edi3, edi4, edi5, mv)

probs_polyreg=fit_polyreg(y,ry,x)

# For each missing value
for (j in 1:n_miss) {
k=row_miss_ext[j]

count=0
reject=1

post <- matrix(probs_polyreg[j,], nrow = 1, ncol = length(probs_polyreg[j,]))

while(reject==1 & count<max_iter_reject) {
count=count+1

# Draw random imputation for extent
un <- rep(runif(1), each = 4)
     draws <- un > apply(post, 1, cumsum)
ext_imp <- 1 + apply(draws, 2, sum)

reject = check_comp_ext(ext_imp,data1$lambda_p[k],data1$cum_lambda_p[k],
                data1$lambda_baseline[k],data1$cum_lambda_baseline[k],
             data1$status[k],data1$age_z[k],data1$men[k],data1$edi2[k],
    data1$edi3[k],data1$edi4[k],data1$edi5[k],
    data1$site[k],data1$mv[k],betas_subst)
}
print(paste("count=",count))
# Save iteration
imputed_ext[j,l]=ext_imp

# Replace extent values in data by newly imputed ones
data1[k,"extent"]=ext_imp
}
write.matrix(imputed_ext,file=paste("v4_imputed_ext",l,".txt",sep=""),sep="\t")
}

# Save completed dataset
data_compl[[imp]]=data1
write.matrix(data1,file=paste("v4_completed_dataset",imp,".txt",sep=""),sep="\t")
}
return(data_compl)
}


# Function to calculate mode of a vector
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Function for sample rejection - Extent
check_comp_ext <- function(extent,lambda_p,cum_lambda_p,lambda_baseline,cum_lambda_baseline,
         d,age_z,men,edi2,edi3,edi4,edi5,site,mv,betas_subst) {

ext2=ext3=ext4=0
ext2=ifelse(extent==2,1,0)
ext3=ifelse(extent==3,1,0)
ext4=ifelse(extent==4,1,0)
```

```
#print(paste("extent=",extent))
#print(paste("ext2=",ext2))
#print(paste("ext3=",ext3))
#print(paste("ext4=",ext4))

if (d==0) {
#print("0")
c_lim = exp(-cum_lambda_p) * exp(-cum_lambda_baseline*
  exp(betas_subst["age_z"]*age_z+betas_subst["men"]*men+betas_subst["edi2"]*edi2+
  betas_subst["edi3"]*edi3+betas_subst["edi4"]*edi4+betas_subst["edi5"]*edi5+
betas_subst["extent2"]*ext2+betas_subst["extent3"]*ext3+betas_subst["extent4"]*ext4+
betas_subst["site"]*site+betas_subst["mv"]*mv ))
u=runif(1)
reject=ifelse(u<=c_lim,0,1)
} else if (d==1) {
#print("1")
A=cum_lambda_baseline*lambda_p/lambda_baseline
B=exp(betas_subst["age_z"]*age_z+betas_subst["men"]*men+betas_subst["edi2"]*edi2+
 betas_subst["edi3"]*edi3+betas_subst["edi4"]*edi4+betas_subst["edi5"]*edi5+
 betas_subst["extent2"]*ext2+betas_subst["extent3"]*ext3+betas_subst["extent4"]*ext4+
 betas_subst["site"]*site+betas_subst["mv"]*mv)
c_lim=((lambda_p+lambda_baseline*B)*exp(-cum_lambda_p-cum_lambda_baseline*B))/
((lambda_baseline/cum_lambda_baseline)*exp(-cum_lambda_p-1+A))
u=runif(1)
reject=ifelse(u<=c_lim,0,1)
}
return(reject)
}

with_mexhaz_smcfcs_app <- function(imputed,m)  {
coef=matrix(NA,ncol=m,nrow=18)
rownames(coef)=c("age_z","men","edi2","edi3","edi4","edi5","extent2","extent3","extent4",
     "age_z_std","men_std","edi2_std","edi3_std","edi4_std",
     "edi5_std","extent2_std","extent3_std","extent4_std")
for (i in 1:m) {
data=imputed[[i]]
fit=mexhaz(formula=Surv(time=survtime_yrs, event=status)~age_z + men +
              edi2 + edi3 + edi4 + edi5 + extent,
      data=data, base="exp.bs", degree=3, knots=c(1), expected="popmort_spec",
        verbose=0, n.gleg=50, fnoptim="optim")

coef["age_z",i]=fit$coefficients["age_z"]
coef["men",i]=fit$coefficients["men"]
coef["edi2",i]=fit$coefficients["edi2"]
coef["edi3",i]=fit$coefficients["edi3"]
coef["edi4",i]=fit$coefficients["edi4"]
coef["edi5",i]=fit$coefficients["edi5"]
coef["extent2",i]=fit$coefficients["extent2"]
coef["extent3",i]=fit$coefficients["extent3"]
coef["extent4",i]=fit$coefficients["extent4"]

coef["age_z_std",i]=fit$std.errors["age_z"]
coef["men_std",i]=fit$std.errors["men"]
coef["edi2_std",i]=fit$std.errors["edi2"]
coef["edi3_std",i]=fit$std.errors["edi3"]
coef["edi4_std",i]=fit$std.errors["edi4"]
coef["edi5_std",i]=fit$std.errors["edi5"]
coef["extent2_std",i]=fit$std.errors["extent2"]
coef["extent3_std",i]=fit$std.errors["extent3"]
coef["extent4_std",i]=fit$std.errors["extent4"]
}
return(coef)
}

combine_mexhaz_app <- function(coef) {
ncoef=(dim(coef)[1])/2
m=dim(coef)[2]
results=matrix(NA,ncol=4,nrow=ncoef)
rownames(results)=c("age_z","men","edi2","edi3","edi4","edi5","extent2","extent3","extent4")
colnames(results)=c("mean","std.err","lower","upper")
for (i in 1:ncoef) {
results[i,1]=mean(coef[i,])
U=mean(coef[(i+ncoef),]^2)
B=1/(m-1)*sum((coef[i,]-mean(coef[i,]))^2)
results[i,2]=sqrt(U+(1+1/m)*B)
# Calculate confidence interval limits
df=floor((m-1)*(1+U/B)^2)
t=qt(0.975,df)
results[i,3]=results[i,1]-t*results[i,2]
results[i,4]=results[i,1]+t*results[i,2]
}
return(results)
}
```