From the
INSTITUTE OF ENVIRONMENTAL MEDICINE
Karolinska Institutet, Stockholm, Sweden

A PERCENTILE APPROACH TO TIME-TO-EVENT OUTCOMES

Andrea Bellavia

Karolinska Institutet

Stockholm 2015

Cover picture by Maria Bellavia
Published by Karolinska Institutet
Printed by E-Print AB 2015

ISBN 978-91-7676-106-9

A PERCENTILE APPROACH TO TIME-TO-EVENT OUTCOMES

THESIS FOR DOCTORAL DEGREE (Ph.D.)

by

Andrea Bellavia

Principal Supervisor:
Associate Professor Nicola Orsini
Karolinska Institutet
Institute of Environmental Medicine
Unit of Nutritional Epidemiology
Unit of Biostatistics

Co-supervisors:
Professor Matteo Bottai
Karolinska Institutet
Institute of Environmental Medicine
Unit of Biostatistics

Professor Alicja Wolk
Karolinska Institutet
Institute of Environmental Medicine
Unit of Nutritional Epidemiology

Opponent:
Professor Nicholas P. Jewell
University of California, Berkeley
School of Public Health
Department of Statistics

Examination Board:
Professor Finn Rasmussen
Karolinska Institutet
Department of Public Health Sciences

Professor Wulf Becker
National Food Agency
Risk-Benefit Assessment Department

Senior Lecturer Mark Clements
Karolinska Institutet
Department of Medical Epidemiology
and Biostatistics

He who has reached the stage where he no longer wonders about anything, merely demonstrates that he has lost the art of reflective reasoning

Max Planck

Abstract

Evaluating survival percentiles is a possible approach for the analysis of time-to-event outcomes that moves the focus from risk to time, as the proportion of events is fixed and the time by which that proportion is achieved is investigated. The development of statistical methods for conditional censored quantiles has opened up the possibility to use this approach in epidemiological studies. The aim of this doctoral thesis was to investigate the advantages of this method in epidemiology, by evaluating survival percentiles in observational studies on lifestyle and mortality, and by extending and further developing the statistical tools for the estimation of conditional survival percentiles.

The percentile approach was used in a large prospective cohort of about 80.000 middle-aged and elderly Swedish men and women, with 15 years of follow-up during which 20% of the study population died. The impact of modifiable lifestyle factors such as fruit and vegetables consumption (Study I), sleep duration and physical activity (Study II), and processed and non-processed red meat consumption (Study III) on time to death were evaluated. Statistical modeling of conditional survival percentiles was conducted using Laplace regression. The epidemiological measure of exposure-outcome association was defined in terms of percentile difference (PD). Quantitative exposures were flexibly modeled using splines to investigate the dose-response shape.

Low fruit and vegetables consumption (Study I) was found to be associated with progressively shorter survival up to 3 years (PD: -37 months; 95% CI: -58, -16) when comparing those who consumed 5 servings/day and those who never consumed fruit and vegetables. Long sleep duration, over 8 hours/day, (Study II) was associated with shorter survival (PD = -20 months; 95% CI: -30, -11) among those with low physical activity, comparing with those with 7 hours of sleeping per day. In Study III, compared with no consumption, higher intake of processed red meat (200 g/d) was associated with shorter survival (PD: -10 months; 95% CI: -18, -3). High and moderate intakes of non-processed red meat were associated with shorter survival only when accompanied by a high intake of processed red meat.

Study IV and Study V introduced novel developments and extensions of the percentile approach. Study IV presented the meaning and evaluation of survival percentiles in those situations where the time variable of interest is attained age at the event rather than follow-up time. This change in the time-scale has important consequences on the definition and interpretation of the survival curve and related percentiles. The study described how to use multivariable Laplace regression models to estimate percentiles of age at death conditioning on age at entry into the study, exposures, and potential confounders. Study V focused on interaction analysis. Interaction can be evaluated on the additive or multiplicative scale, but its assessment in prospective studies is commonly limited to the multiplicative scenario. In this study the advantages of using a percentile approach in interaction analysis were presented. A measure of interaction in terms of time was introduced and how Laplace regression can be used to estimate a measure of interaction on the additive scale was described.

Evaluating survival percentiles provides an intuitive and flexible approach for the analysis of time-to-event outcomes. With this method, results from prospective studies can be presented in terms of differences in survival time, facilitating both interpretation and communication of scientific findings. The introduction of a statistical technique to estimate conditional survival percentiles has substantially enriched its potentialities and eased its application in epidemiological research. The percentile approach should be considered as a possible complement to classical approaches and its use should be widespread.

List of publications

I. **Bellavia A.**, Larsson SC., Bottai M., Wolk A., Orsini N.
Fruit and vegetable consumption and all-cause mortality: a dose-response analysis.
American Journal of Clinical Nutrition. 2013 Aug;98(2):454-9

II. **Bellavia A.**, Åkerstedt A., Bottai M., Wolk A., Orsini N.
Sleep duration and survival percentiles across categories of physical activity.
American Journal of Epidemiology. 2014 Feb 15;179(4):484-91

III. **Bellavia A.**, Larsson SC., Bottai M., Wolk A., Orsini N.
Differences in survival associated with processed and with nonprocessed red meat consumption.
American Journal of Clinical Nutrition. 2014 Sep;100(3):924-9

IV. **Bellavia A.**, Discacciati A., Bottai M., Wolk A., Orsini N.
Using Laplace regression to model and predict percentiles of age at death when age is the primary time scale.
American Journal of Epidemiology. 2015 Jun;182(3):271-277

V. **Bellavia A.**, Bottai M., Orsini N.
Evaluating additive interaction using survival percentiles.
Epidemiology. 2016. In press

The articles will be referred to in the text by their Roman numerals, and are reproduced in full at the end of the text.

Related publications

- **Bellavia A.**, Bottai M., Wolk A., Orsini N.
  Physical activity and mortality in a prospective cohort of middle-aged and elderly men - a time perspective.
  International Journal of Behavioral Nutrition and Physical Activity. 2013 Aug 8;10:94

- **Bellavia A.**, Bottai M., Wolk A., Orsini N.
  Alcohol consumption and mortality: a dose-response analysis in terms of time.
  Annals of Epidemiology. 2014 Apr;24(4):291-6

- **Bellavia A.**, Bottai M., Discacciati A., Orsini N.
  Adjusted survival curves with multivariable laplace regression.
  Epidemiology. 2015 Mar;26(2):e17-8.

- Rahman I., **Bellavia A.**, Wolk A., Orsini N.
  Physical activity and heart failure risk in a prospective study of men.
  JACC: Heart Failure. 2015; 3(9), 681-687

- **Bellavia A.**, Wolk A., Orsini N.
  Differences in age at death according to smoking and age at menopause.
  Menopause. 2015 Jul 31. [Epub ahead of print]

- Discacciati A., **Bellavia A.**, Orsini N., Greenland S.
  On the interpretation of risk and rate advancement periods.
  International Journal of Epidemiology. 2016. Conditionally accepted

- **Bellavia A.**, Teknonidis TG., Orsini N., Wolk A., Larsson SC.
  Quantifying the benefits of mediterranean diet in terms of survival.
  European Journal of Epidemiology. 2016. Conditionally accepted

List of abbreviations

AL	Asymmetric Laplace
AFT	Accelerated Failure Time
BMI	Body Mass Index
CI	Confidence Interval
COSM	Cohort of Swedish Men
CVD	Cardiovascular Disease
FV	Fruit and Vegetables
HR	Hazard Ratio
MET	Metabolic Equivalent
PA	Physical Activity
PD	Percentile Difference
PH	Proportionality of the Hazard
RERI	Relative Risk due to Interaction
SD	Standard Deviation
SMC	Swedish Mammography Cohort

Contents

1. Introduction

Epidemiology, by evaluating the determinants and distribution of the diseases at the population level, plays a crucial role among the public health sciences. In the last decades, epidemiological studies have provided a substantial contribution in identifying modifiable lifestyle and behavioral risk factors for non-communicable diseases, which are responsible for a large fraction of deaths worldwide. Results from epidemiological studies have been used to plan interventions, to build policy decisions, and to make public health recommendations.

To assess the public health impact of modifiable risk factors, is critical to understand how these affect the time involved in developing the disease of interest and the overall survival of the population. To take into account the time dimension of the disease, epidemiological studies are commonly designed in a prospective setting. The prospective cohort study, in particular, assures that all possible risk factors are assessed before the subjects enrolled into the study develop the disease of interest, and it is designed to investigate the risk of the event of interest and the time until this is observed.

The statistical evaluation of time-to-event outcomes, due to peculiar characteristics, can not be carried out with classical statistical methods. The most popular approach to summarize time-to-event data is to display hazard ratios of the event according to levels of the risk factor of interest calculated over an observational period. This method, while providing considerable advantages, is also subject to some recognized limitations, which often prevent results to be translated into meaningful public health messages. Over the recent years, different alternative approaches to complement current methodologies, such as the evaluation of survival percentiles, have been proposed, but the hazard ratio solidly remains the most common tool adopted to summarize results from prospective cohort studies. One explanation for the popularity of hazard ratios in epidemiology is certainly the need of adjusting for confounders, for which statistical models are required. Although alternative measures to summarize time-to-event data have been proposed, their introduction has not always been complemented by the development of regression methods for their estimation.

This thesis aims to shorten this gap in the epidemiological literature, by taking advantage of recent methodological developments for the analysis of time-to-event outcomes to provide a regression-based framework for the estimation of conditional survival percentiles in epidemiological studies.

2. Background

2.1 Lifestyle and health

Around 38 million deaths per year can be attributed to noncommunicable diseases, also known as chronic diseases (World Health Organization, 2009). Modifiable lifestyle factors such as dietary factors, tobacco use, harmful use of alcohol, and physical inactivity, are recognized risk factors for the development of major chronic diseases such as cardiovascular diseases, cancer, and diabetes. A recent report from the Global Burden of Diseases has showed (Figure 2.1) that most of the worldwide number of deaths and of life-years lost can be attributed to lifestyle and behavioral modifiable risk factors, especially in high-income countries (Global Burden of Disease, 2015). In this scenario it is evident the prominent role that prevention and primary care policy could play. Changes in the lifestyle, especially in terms of improving diet, increasing levels of physical activity, and smoking cessation, have been recognized as a primary tool to reduce the worldwide burden of deaths and diseases, and different public health efforts have been made to promote adopting a healthy lifestyle (World Health Organization, 2004).

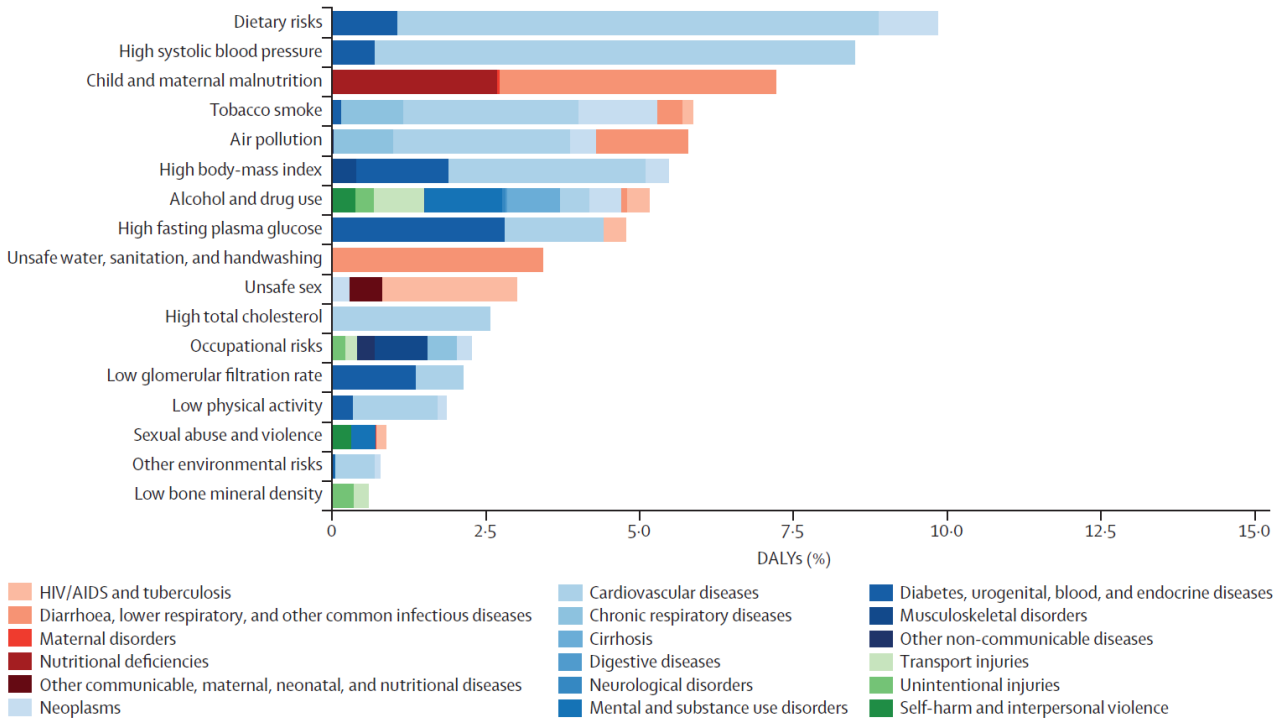


Figure 2.1: Global disability-adjusted life-years (DALYs) attributed to level 2 risk factors in 2013 for both sexes combined. Reproduced with permission from *The Lancet*, 2015.

Epidemiology, by studying the patterns and causes of diseases in different populations, has played a leading role in understanding and quantifying the global effects of diet and unhealthy behaviors, and a large number of observational studies have contributed in understanding the importance of diet and healthy lifestyle behaviors in preventing diseases. Epidemiological studies commonly evaluate the association between one *exposure* of interest, or the *interaction* between two or more exposures, and the development of a health-related outcome, presenting results in terms of relative changes (e.g. risk/rate).

To assess the health benefits of nutritional and behavioral factors, and to promote lifestyle recommendations, an essential information that needs to be investigated is how these factors influence the overall survival experience of the population (Michels, 2003; Rothman et al., 2008). However, despite the recognized relevance of evaluating the impact of risk factors on overall mortality, observational studies often prefer to focus on assessing the incidence of a given disease, or mortality from a specific cause. Although these studies certainly provide insightful information, these are not enough to fully depict the health-effects of adhering to recommended lifestyle factors.

There are different behavioral and dietary factors that are known to be either protective or harmful for specific diseases, but whose association with mortality from all-causes is not established. Remarkable examples of controversial risk factors are the consumption of fruit and vegetables, the daily amount of sleeping hours, and the intake of red meat.

2.1.1 Fruit and vegetables consumption

A diet low in fruit and vegetables (FV) has been associated with higher risk of developing major chronic diseases such as CVD (Bazzano et al., 2003; Dauchet et al., 2006) and cancer (Boffetta et al., 2010), and current dietary guidelines recommend a daily consumption of 5 or more servings of FV (World Health Organization, 2002). Nevertheless, the association between FV consumption and overall mortality has been only investigated by few studies with inconsistent findings (Rissanen et al., 2003; Steffen et al., 2003; Hung et al., 2004; Genkinger et al., 2004; Agudo et al., 2007), and most of the current knowledge on the association between FV and overall mortality comes from relatively small studies that examined the intake of total serum carotenoids as a marker of FV (Waart et al., 2001; Ray et al., 2006; Lauretani et al., 2008; Shardell et al., 2011; Nicklett et al., 2012).

2.1.2 Sleep duration

Different studies have suggested that the association between sleep duration and mortality is described by a U-shape. A daily sleep duration of 7 hours has been indicated to provide the largest benefits, while both longer and shorter sleep duration are associated with the development of adverse health conditions (Kripke et al., 2002; Gangwisch et al., 2008; Hublin et al., 2007; Cappuccio et al., 2010). While biological mechanisms to explain the negative effects of short sleep duration have been theorized, some concerns have been raised on the negative effects of long sleep duration and the underlying biological mechanisms (Knutson and Turek, 2006; Stamatakis and Punjabi, 2007). Among possible factors that could confound or mediate the U-shaped association between sleep duration and mortality, low physical activity was found to be strongly associated with long sleep duration, and some authors suggested that long sleep may actually represent an epiphenomenon of comorbidity with low physical activity (Stranges et al., 2008). This hypothesis, however, has never been tested.

2.1.3 Red meat consumption

High red meat consumption has been associated with an increased risk of various chronic diseases (Larsson et al., 2006; Sinha et al., 2009; Chan et al., 2011; Micha et al., 2012; Kaluza et al., 2012; Pan et al., 2012) and higher mortality (Larsson and Orsini, 2013). The total intake of red meat is commonly evaluated by dividing the exposure into processed and non-processed (fresh) meat consumption, which might have a different biological impact on health. While the unhealthy effects of processed meat consumption are largely established (Marmot et al., 2007), the role of non-processed meat is still unclear. Recent studies have suggested that fresh meat consumption might be associated with higher mortality (Sinha et al., 2009; Pan et al., 2012; Rohrmann et al., 2013), while other have reported no association (Takata et al., 2013; Larsson and Orsini, 2013; Kappeler et al., 2013). No studies have evaluated the joint association of processed and non-processed meat consumption with mortality from all-causes, which would allow understanding the interrelationship of the two components.

An additional common limitation of observational studies evaluating all-cause mortality is the tendency of presenting results only in relative terms. However, when the probability of the outcome is 1 (*certain events*) a relevant information, if not the most relevant, is the timing to the event, as the interest lies in understanding the extent by which possible risk factors delay or anticipate the occurrence of the event. Few studies have so far attempted to quantify the impact of lifestyle and behavioral factors on health in terms of survival time. The next section will present the methodologies commonly adopted to analyze time-to-event outcomes and introduce possible alternatives that allow to summarize research findings in terms of time.

2.2 Statistical methods for time-to-event outcomes

A common study design adopted in epidemiology is the *prospective cohort study*, where exposures are assessed at the beginning of the observational period (*follow-up*) and participants are followed until they experience the event of interest (Rothman et al., 2008).

When data are collected prospectively, researchers are interested in investigating both the rate/probability of the event of interest over time, and the time until this event occurs, conditioned on independent predictors. Survival analysis is the statistical branch dealing with methods to evaluate time-to-event variables. Different peculiar features make survival methods unique within the field of medical statistics. First, as the name *time-to-event* suggests, there are two equally interesting quantities: the event (commonly a binary variable D), and the time T to the development of the event (commonly a continuous variable[1]). The second important property is that the time variable T is always positive and commonly skewed, with large numbers of events observed either at the beginning or at the end of follow-up. For this reason, standard statistical procedures that assume normality of distributions, such as the classical linear regression, are of limited use in survival analysis. Another distinctive feature of survival data is that they commonly include individuals for whom the event D is not observed during follow-up. In prospective cohort studies this is generally caused by *right-censoring*, which occurs when individuals are included into the study but have not experienced the event of interest at the end of the observational period.

Two informative quantitative tools to summarize and describe the distribution of events over time are the survival function $S(t)$, and the hazard function $h(t)$,[2] which are depicted in Figure 2.2.

The hazard function represents the instantaneous failure rate and is defined as the probability that the event occurs in a specific time interval, given that the subject has survived to the beginning of the interval, and divided by the time interval:

$$h(t) = \lim_{\Delta \to 0} \frac{Pr(t + \Delta t > T > t | T > t)}{\Delta t} \tag{2.1}$$

The survival function is defined, at any time t, as the probability of surviving beyond t:

$$S(t) = Pr(T > t) \tag{2.2}$$

[1]Sometimes, for example when data are grouped into observations, T may be treated as a discrete random variable.

[2]These two functions are linked by the mathematical relation $S(t) = \exp[-\int_0^t h(u)du]$.

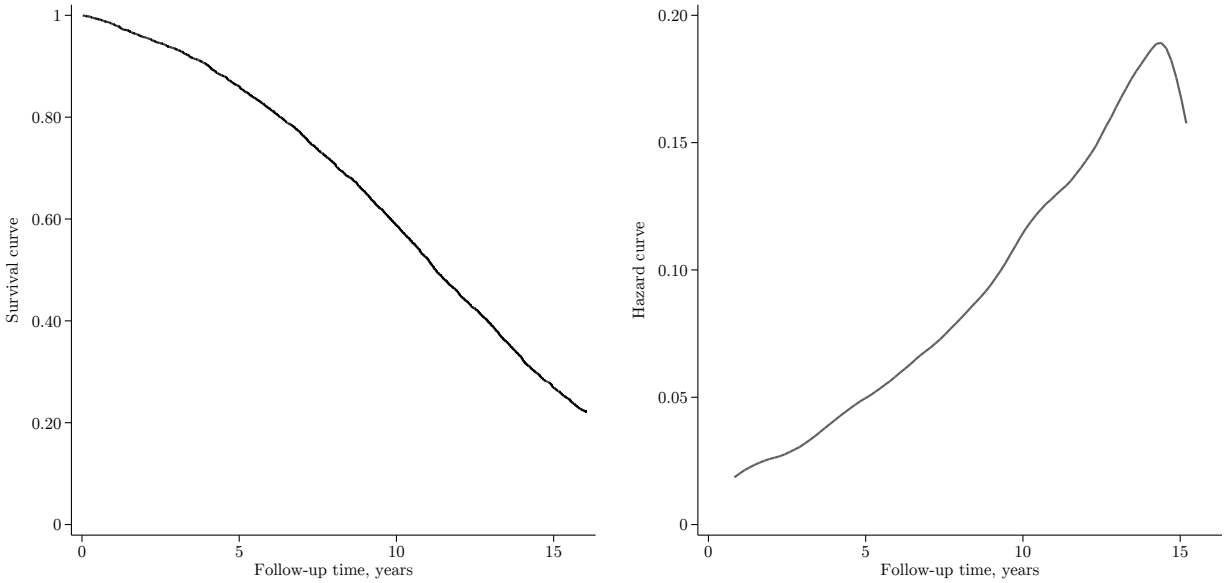


Figure 2.2: Survival and hazard curves in a study population during 16 years of follow-up.

The survival function is constrained between 1 and 0 and is a monotonic non-increasing function of time. On the other hand, the hazard function does not represent a probability and has no upper bounds. It is continuous and always non-negative, but differently from the survival function is free to start at any value and to move in any direction over time.

The hazard function, by providing a dynamic description of how the instantaneous risk of failing varies, gives little emphasis on the time component, and is an useful tool when one wants to estimate the *rate* of the event over time. The survival function, which is a monotone non-increasing function of time, is an optimal tool to combine information on the *risk* of the event, and the *time* by which this risk is achieved (Andersen et al., 2012). Statistical methods to investigate time-to-event outcomes have been mainly developed as estimators of these two functions. In the following subsections, the Kaplan-Meier estimator and the Cox regression, which are the two most popular methods used in medical research to evaluate survival data, will be presented.[3]

[3]The introductory papers of these two methods are listed among the 100 most-cited papers of all time and occupy the first two positions in the field of statistics (Van Noorden et al., 2014). Among the other methods for time-to-event outcomes, noteworthy are the Nelson-Aalen estimator for the cumulative hazard function (defined as $H(t) = \int_0^t h(t)dt$) (Nelson, 1972), and the parametric estimation of the survival and hazard curves through means of accelerated-failure times models (AFT) (Wei, 1992).

2.2.1 Kaplan-Meier estimator

Kaplan-Meier is a non-parametric method to estimate the survival curve $S(t)$, also known as the *product limit* estimate of $S(t)$ (Kaplan and Meier, 1958). The estimator is given by

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) \tag{2.3}$$

where n_j is the number of individuals at risk at time t_j and d_j is the number of failures at time t_j.

The Kaplan-Meier estimator is generally used to investigate and present the observed survival experience of the entire population, or stratified by different subpopulations. In the clinical trials literature Kaplan-Meier is presented as the primary way to analyze time-to-event data (Friedman et al., 2010). In epidemiological studies, however, results commonly need to be adjusted for a variety of possible confounders, to understand the effects of independent risk factors on the survival experience of the study population. Methods to calculate survival curves while adjusting for possible confounders (*adjusted survival curves*) have been developed (Nieto and Coresh, 1996; Ghali et al., 2001; Austin and Schuster, 2014), but these are rarely applied when the exposures-outcome associations need to be further adjusted for other relevant factors. The Kaplan-Meier method, eventually, is generally included in observational studies as a preliminary analysis.

2.2.2 Cox proportional-hazard model

From its introduction (Cox, 1972), the Cox proportional hazard (PH) model has rapidly grown to become the most-widely used method to analyze survival data. The method estimates hazard ratios (HRs), which are measures of the extent to which a covariate increases or decreases the rate of the event, assuming that the effect of the included covariates is multiplicative.

The basic assumption of the Cox model is that the hazards of two populations (e.g. exposed and unexposed) are proportional over time, implying that the estimated parameters are constant with respect of the follow-up time. A Cox model takes the form

$$h_i(t|\mathbf{x}_i) = h_0(t) exp(\mathbf{x}_i^T \boldsymbol{\beta}) \tag{2.4}$$

The hazard function $h_i(t|\mathbf{x}_i)$ is parametrically modeled as a function of the covari-

ates $\mathbf{x}_i$, while no shape needs to be assumed for the baseline hazard h_0. For this reason the model is often referred to as *semi-parametric*.

The Cox model has been largely explored and different books presenting the method in details have been published (e.g. Hosmer et al. (2011); Kleinbaum and Klein (2012)). Extensions to cover situations of time-varying effects and other methodological settings have been also investigated, and good introductions to possible extensions of the model can be read in Therneau and Grambsch (2000); Royston and Lambert (2011); Andersen et al. (2012).

2.2.3 Definition of the time variable

The time variable T has been so far defined as the time between entering into the study and either experiencing the event of interest or censoring. In prospective studies, however, other possible definitions of time can be chosen (Kleinbaum and Klein, 2012). A common alternative to follow-up time is to focus on the age of the participants, defining the time variable on the age scale. Instead of entering the study at the same point in time, each individual is assumed to enter at his/her own baseline age, and is followed until the age at which he/she experiences the event of interest or is censored.

The consequences of changing time-scale on the hazard function, and how this affect the analyses when Cox regression is used, have been widely investigated (Lamarca et al., 1998; Cheung et al., 2003; Thiébaut and Bénichou, 2004; Westreich et al., 2010; Cologne et al., 2012). When data are analyzed with Cox regression, it is becoming increasingly common to use attained age at the event of interest as primary time-scale, as this choice allows a more flexible modeling of age in the non-parametric part of the model (Kom et al., 1997; Liestol and Andersen, 2002; Cheung et al., 2003; Thiébaut and Bénichou, 2004; Westreich et al., 2010; Cologne et al., 2012).

The consequences of changing time-scale on the survival function and its estimators, on the other hand, have been seldom investigated.

2.2.4 Hazard ratios

Different advantages make the estimation of HRs through Cox PH regression an appealing choice for researchers dealing with time-to-event outcomes (Kleinbaum and Klein, 2012). Nevertheless, HRs are subject to some established limitations, and relying on HRs alone to present statistical associations may prevent a correct interpretation and

translation of scientific findings.

First, when HRs are derived with statistical models that take advantage of the PH assumption,[4] but the assumption is violated, they lose their meaning in summarizing exposure-outcome associations, as they do not represent the average HR of the true HRs over follow-up (Kalbfleisch and Prentice, 1981). In addition, using single estimates to summarize situations in which the HR is varying over time would eventually produce wrong and misleading conclusions. As an illustrative example, Figure 2.3 depicts a possible scenario in which the HR is <1 at the beginning of follow-up, and >1 after 6 years of follow-up; in such situations using a single HR to present the association would not correctly summarize the information contained in the data.

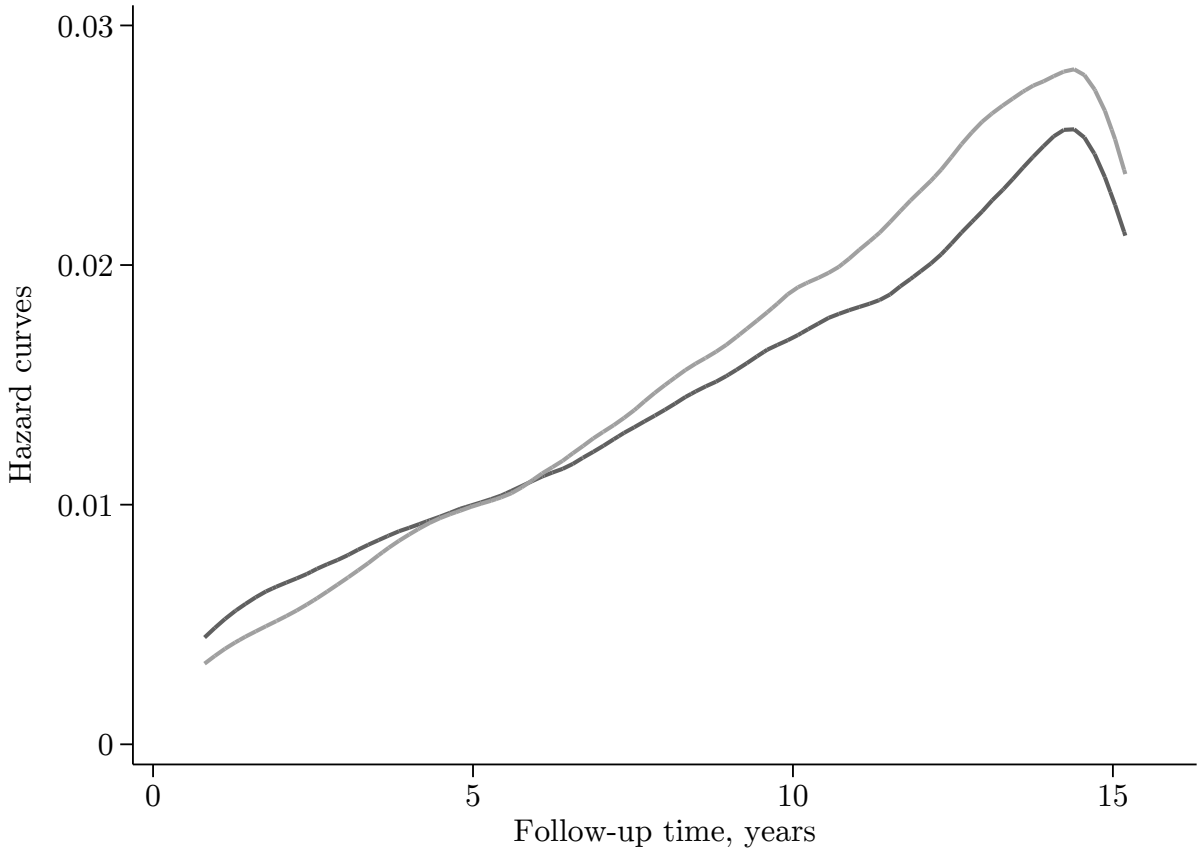


Figure 2.3: Hazard curves in two sub-populations during 16 years of follow-up.

[4]The assumption of PH is shared by parametric models for the survival and the hazard function.

HRs are also subject to two important limitations that make challenging their interpretation and their translation into meaningful public health measures. Such limitations are built-in the measure, and persist even in those situation when the PH assumption is plausible. First, HRs are unit-less measures, and the lack of information on the background risk prevents results to be translated into transparent public health messages (Uno et al., 2014). The baseline hazard is a clinically meaningful tool to understand the course of the disease of interest, and is the ground against which HRs are estimated (Royston and Lambert, 2011). For example, an HR equal to 2 implies that, compared to a reference group, the average hazard over time of a subgroup of interest is twice higher, but does not provide any information on the absolute background risk. We can interpret the result in terms of relative change, but we don't know what is this change compared to. It has been pointed out that this lack on information on the absolute risk makes HR's interpretation difficult for laymen, and not sufficient to support clinical decisions (Lytsy et al., 2012; Uno et al., 2014). It was shown, for example, that relative measures such as the HR, when presented to doctors, exaggerate understanding of the effect. (Naylor et al., 1992; Forrow et al., 1992; Bobbio et al., 1994; Hux and Naylor, 1995). The importance of clear and easy-to-understand summary measures is extremely relevant in epidemiology, as this field plays an important role in the process of translating scientific discoveries into population health impact (Khoury et al., 2010). This drawback of the HRs is of particular importance when studies are investigating certain events such as all-cause mortality, as the most relevant question is not *if* the event occurs, but *when* the event occurs.

A second major limitation of HRs is that they are dependent on the length of follow-up, because the average HR ignores the distribution of events during the observational period (Hernán, 2010; Uno et al., 2014). Two studies evaluating the same population and investigating the same exposure-outcome association, but with different follow-up lengths, could result in different estimates of the HRs regardless the sample size. Conversely, the same HR could be obtained in two different studies but conducted over a different follow-up time.[5]

[5]A persuasive example is given in Hernán (2010), where the HR of interest would have been 1.8 if the study had been halted after 1 year, 1.7 after 2 years, and 1.2 after 5 years, despite the PH assumption being not violated.

2.2.5 Alternative measures

When the hazard function contains relevant information to a specific study, and one is satisfied with providing associations in relative terms, estimating HRs via Cox regression represents the best option to analyze time-to-event data, after careful evaluation of the PH assumption validity. However, when the discussed limitations might represent a threat for the analysis, interpretation, and translation of the results, researchers should think of complementing analyses with methods that focus on the survival curve, summarizing their findings with measures that directly involve time (Hernán, 2010; Uno et al., 2014).

The most intuitive option to summarize information included in the survival curve would be to estimate the life expectancy of the study population (*mean survival*), but due to the censoring mechanism this can not be achieved unless assumptions are made.[6] However, different other approaches are available, and measures of association in the metric of time have been proposed. Among the recommended methods, calculation of t-years survival rates, and restricted mean survival represent two appealing options (Uno et al., 2015). Another possible approach is to focus on the median survival and to present differences in median survival as measures of associations (Friedman et al., 2010; Uno et al., 2014). This approach can be extended to any other percentile of the survival distribution, offering additional insights and important advantages. This method, commonly referred to as *evaluating survival percentiles*, is the main subject of this doctoral thesis and will be thoroughly discussed in the next section. Following sections (2.4 and 2.5) will then explore the use of regression methods to model survival percentiles.

[6]Estimation of life expectancy involves computing the integral of the survival curve, which corresponds to calculating the area under the survival curve. This area, however, will result equal to ∞ if the survival curve does not go down to 0.

2.3 Survival percentiles

In a given study population, where individuals are followed-up over time to investigate the development of an event of interest, survival percentiles[7] can be defined as the time points by which specific proportions of participants have experienced the event. Formally, the pth survival percentile is the time t by which $p\%$ of the study population have experienced the event of interest, while $(100 - p)\%$ have not. For example, the time by which half of the population has experienced the event of interest is defined as 50th survival percentile, or *median survival*.

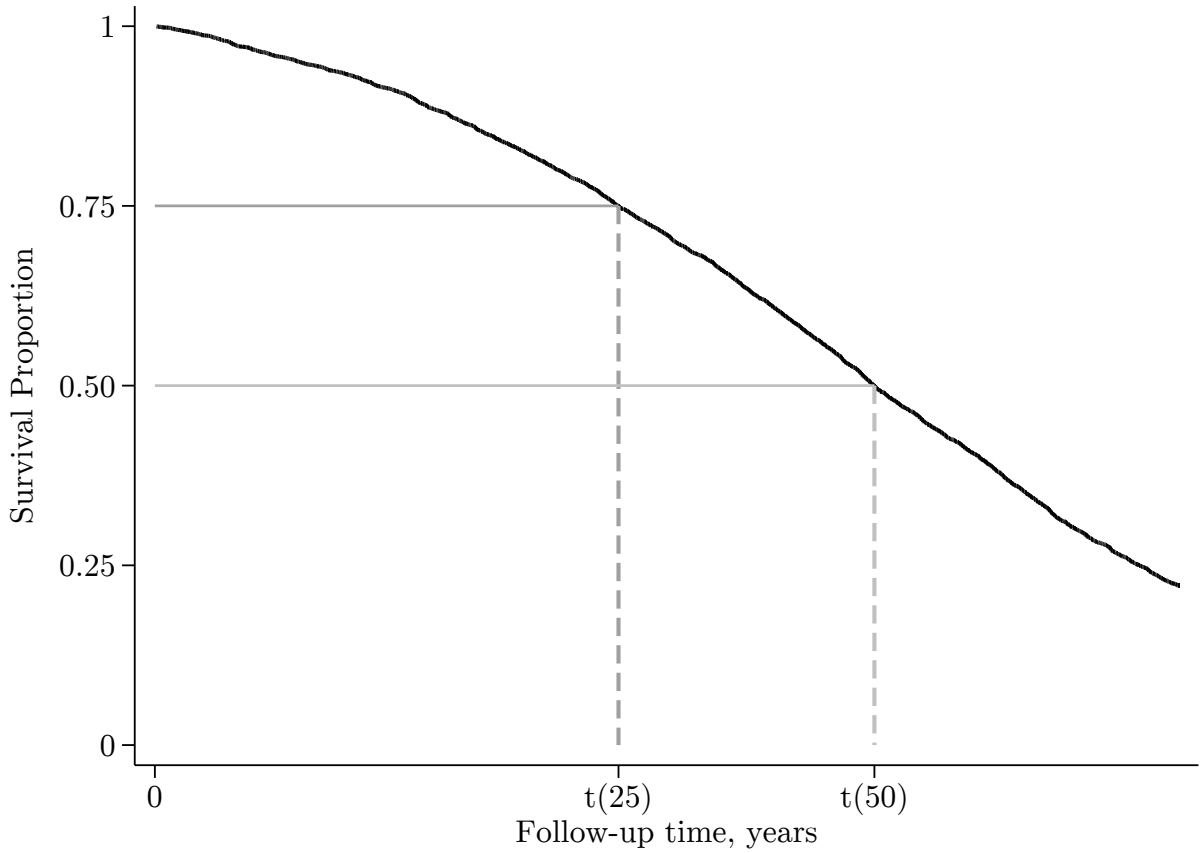


Figure 2.4: Summary of observed survival percentiles: the 25th and 50th percentiles are depicted.

The survival curve, as defined in (2.2), described the probability of the event as a function of time. Survival percentiles are defined by changing the perspective, describing time T as a function of the probability of the event τ. Formally, the quantile func-

[7]Throughout the thesis percentiles and quantiles will be often used interchangeably, as they represent the same quantity. Percentiles (p) are defined within the range 1 to 99, while quantiles (τ) are defined within the range 0.1 to 0.99.

tion $Q(\tau)$ is related to the cumulative distribution function $F(t)$, where $S(t) = 1-F(t)$, by the relationship:

$$F(Q_T(\tau)) = P(T \leq Q_T(\tau)) = \tau \tag{2.5}$$

There is an univocal correspondence between the quantile and the survival function. When T is continuous, $Q(\tau) = t$ only if $F(t) = \tau$, that is, the quantile function is the minimum value of t below which a randomly selected individual from the population will fall $(100 \cdot \tau)\%$ of the times.

In the survival curve presented in Figure 2.4, the 25th and 50th survival percentiles are depicted ($t(25)$ and $t(50)$, respectively). By the time $t(25)$ 25% of the study participants have experienced the event of interest, while 75% have not. By $t(50)$ the event has been observed in half of the population.

Due to the presence of right censoring, the event might not be observed for all participants. However, this does not represent an obstacle for the calculation of survival percentiles, but just limits the evaluation to the range of observed percentiles (1-75 in Figure 2.4). It is also interesting to note that the eventual inclusion of additional years of follow-up would not influence the estimates of previously observed percentiles, but would only possibly increase the number of percentiles that is possible to quantify.

2.3.1 Measures of association: percentile differences

When evaluating survival percentiles a straightforward measure to quantify the variability between two or more exposure groups is to calculate differences in survival percentiles, as depicted in Figure 2.5. The figure represents the survival experience of two populations of participants. Given a fixed proportion of events p, such as the median ($p = 50$), survival percentiles are identified by calculating the time points by which this proportion is achieved in the two different groups (exposed vs unexposed). In the figure, t_0 is the time by which 50% of unexposed individuals (grey line) have experienced the event, while t_1 is the time by which the same fraction of events has been attained by exposed participants (black line).

The exposure effect $t_1 - t_0$ is the difference in time by which participants in the two exposure groups experience the same outcome probability, and can be referred to as difference in the pth survival percentiles (PD_p=pth percentile difference). When $p = 50$, as depicted in Figure 2.5, the measure represents the difference in median survival. PDs are absolute measures of association directly expressed in the unit of time (i.e. days, months, years).

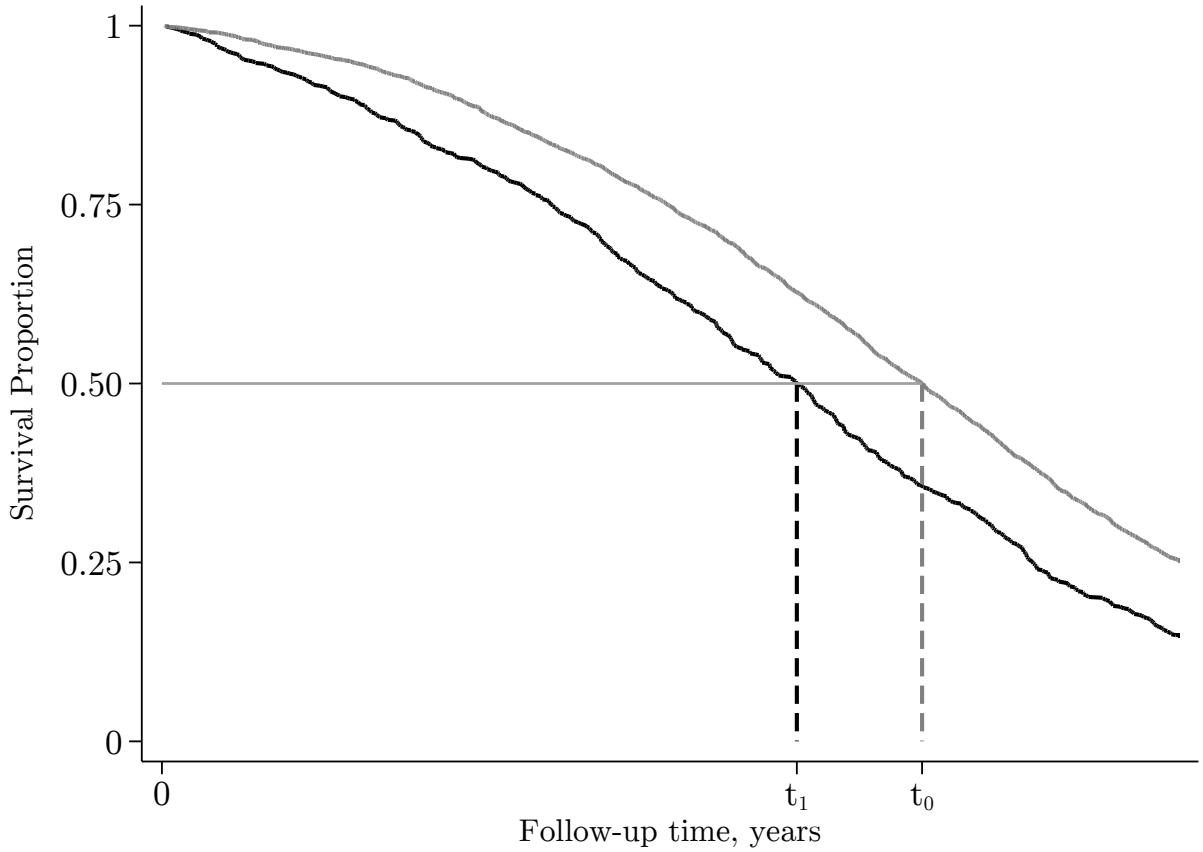


Figure 2.5: Survival percentiles in two sub-populations. The pth percentile is indicated by the horizontal line, and the corresponding survival percentiles are displayed on the x-axis.

PDs can be calculated for any observed percentile. Although presenting one single PD, such as the difference in median survival, could be enough to provide a meaningful summary measure in the metric of time, computing and displaying PDs for all observed percentiles would give additional insights, as it would allow to understand how the association of interest is changing according to the proportion of events observed over time. If one wants to present all observed PDs, a preferable option is to plot PDs as a function of the corresponding proportion of events. Figure 2.6 displays PDs estimated from Figure 2.5 and plotted as a function of p. PDs can be calculated as long as exposed and unexposed participants experience the same risk of the event.

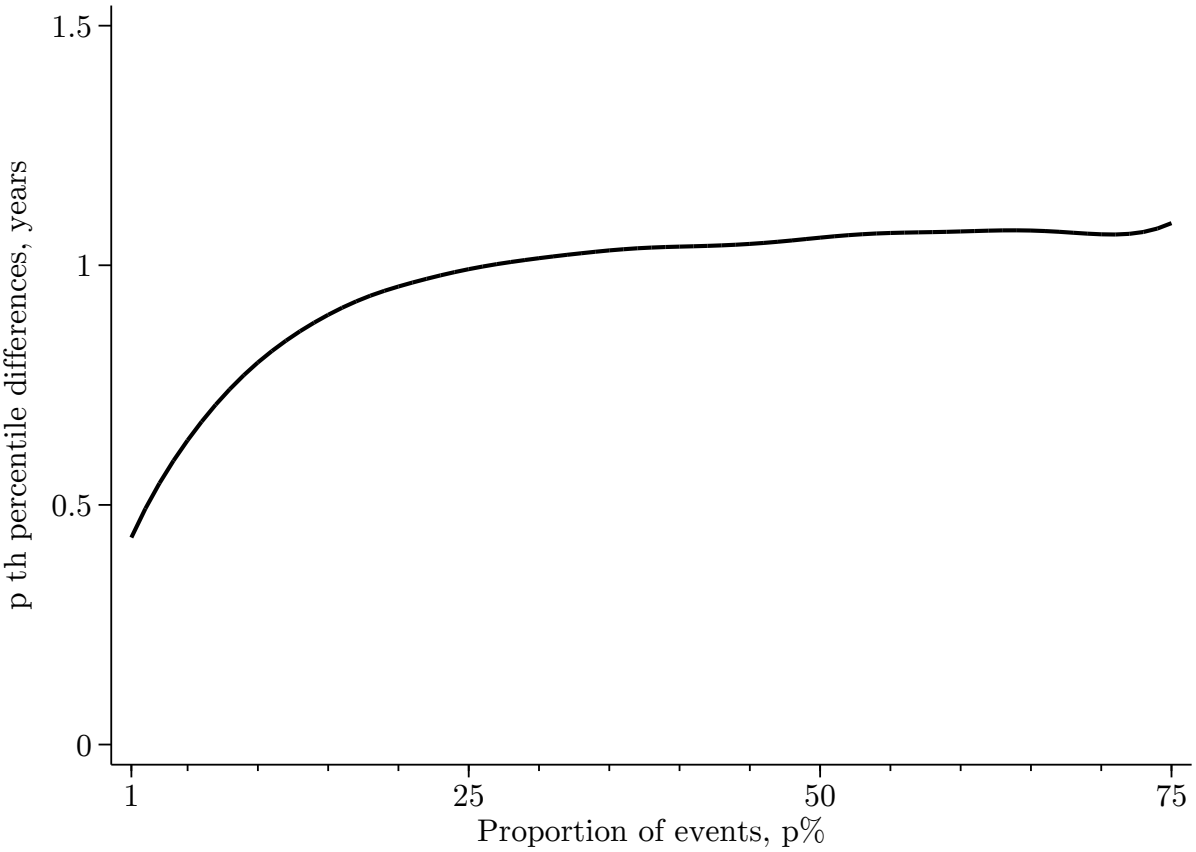


Figure 2.6: Percentile differences as a function of the corresponding fraction of events.

2.3.2 Statistical methods for survival percentiles

The Kaplan-Meier estimator for the survival function can be used to indirectly estimate survival percentiles and differences in survival percentiles (Altman, 1990). Median survival, for example can be derived by estimating $S(t)$ with the Kaplan-Meier product limit estimator and by finding the lowest point t at which $S(t) \leq 0.50$. Confidence intervals for survival percentiles and for PDs may be derived numerically (Brookmeyer and Crowley, 1982)[8] or can be constructed via bootstrap (Efron and Tibshirani, 1994).

In observational studies, however, the frequent need to adjust for potential confounders of the exposure-outcome association of interest, or to assess interactions between exposures in predicting the outcome, require the use of multivariable regression models to address the research question (Rothman et al., 2008). One possibility is to use the relationship between $h(t)$ and $S(t)$ to derive survival percentiles after estimating an hazard-based regression model, but this non-straightforward option might

[8]Formal description on how to calculate confidence intervals for survival percentiles can be found at page 94 of Lawless (2011).

result challenging for many researchers, especially in terms of deriving confidence intervals (Lai and Su, 2006). The absence of established regression-based frameworks to estimate informative measures of associations such as PDs might largely explain their limited application in epidemiological research, and the consistent lack of interest that epidemiologists show towards additional measures different then the HR.

Survival percentiles are formally defined as the quantiles of a time variable. In the last 15 years, statistical methods to evaluate conditional quantiles of possibly censored outcomes have been developed, as extensions to the classical quantile regression approach, introduced in the statistical literature in the late ´70s. These novel statistical methods offer unique modeling advantages, and have opened up the possibility to establish a regression-based approach for the estimation of conditional survival percentiles.

2.4 Quantile regression

2.4.1 Definition

Quantile regression is rapidly growing as an alternative to the classical regression approaches. Ordinary least-squares regression can estimate the mean of a response variable conditional on one or more covariates. Distributions, however, may differ with respect to other aspects than the mean alone, and focusing only on the mean may miss important aspects of the association between outcome and covariates. Quantile regression can estimate the conditional quantiles of a response variable, thus providing a complete view of the effect of explanatory variables on the location, scale, and shape of the distribution of the outcome.

Figure 2.7, which shows the relationship between a continuous exposure and the quantiles of a response variable, is a simple example that illustrates how quantile regression works. The line representing the 50th percentile (median) leaves half of the observations above, and half below, at any given value of the covariate X. In the same way we can define all the other regression lines; regression on the 95th percentile, for instance, leaves 5% of the observations above and 95% below. In this example, the effect of X on Y is different over the distribution of the outcome. While no effects can be detected at lower percentiles, and on the median, higher percentiles of Y are influenced by changes in the predictor X.

A first prototype of median regression was developed by the Jesuit priest Rogerius Boscovich, who in 1757, when studying the ellipticity of the earth, introduced the idea of minimizing the sum of absolute residuals to resolve what we would now address as a regression problem (Boscovich, 1757). The work of Boscovich was investigated and further expanded by Laplace and Edgeworth. However, the computational difficulty of the method was found extremely challenging, and was not until recent years that a formalization of quantile regression was made possible. The early origins of median regression, which is older than mean regression by half a century,[9] is not surprising, as the median is the most immediate and intuitive measure to evaluate the central tendency of a distribution.

The regression framework for the median of a response variable was finally proposed and formalized by Wagner (1959), and extended to other quantiles by Koenker and Bassett (1978), giving birth to the modern quantile regression. Koenker's work on the topic have extended for a long time-span, and his book on quantile regression

[9] The origins of linear regression are commonly dated to 1805, when Legendre introduced the least squares.

is the best available introduction to the subject (Koenker, 2005). A good introduction to quantile regression for epidemiologists can be found in Beyerlein (2014), which discusses the advantages of the method and encourages a wider use.

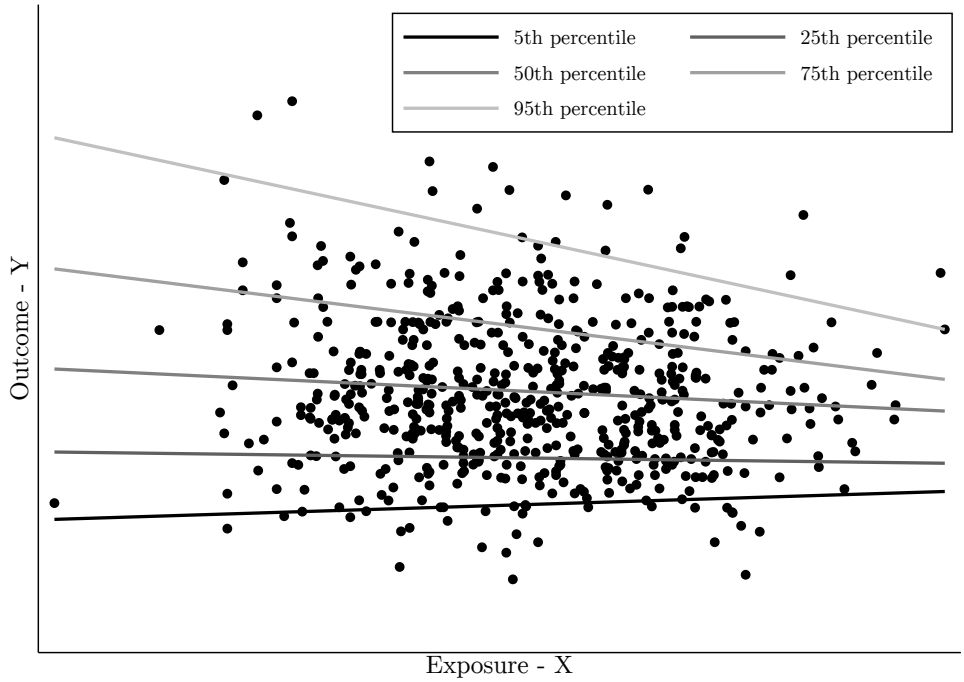


Figure 2.7: Quantile regression lines between a continuous exposure X and a continuous outcome Y.

2.4.2 Estimation

In a quantile regression model, the quantiles of a response variable are modeled as a function of a set of covariates. Following the notation introduced by Koenker we can define, given a quantile τ, a response variable Y_i, and a set of covariates $\mathbf{x}_i$, a linear model for the conditional τth quantile as:

$$Q_{y_i}(\tau|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) \tag{2.6}$$

where $0 < \tau < 1$

Quantile estimation can be seen as an optimization problem. As in the classical linear regression model, where the mean is the solution to the problem of minimizing a sum of squared residuals, the median is the solution to the problem of minimizing a sum of absolute residuals. The quantile-regression distance function can be written as

$$d_\tau(\mathbf{y}, q) = \sum_{i=1}^{n} \rho_\tau(y_i - q) \tag{2.7}$$

where ρ_τ is a weight that, depending on τ, defines all quantiles estimation. When $\tau = 0.5$, the point q where the distance function is minimized is the median of the distribution. The function (2.7) is piecewise linear and continuous, and it is differentiable except at the points in which residuals are equal to 0. This non-differentiability at 0 is the main reason making quantiles estimation more challenging than estimating the mean.[10] Estimation is performed through an iterative algorithm that can be reformulated as a linear programming problem, with estimation of coefficients for each quantile based on the weighted data of the whole sample (Koenker and Bassett, 1978).

The Asymmetric Laplace distribution

A common option to estimate quantiles is to assume that the outcome Y_i follows an asymmetric laplace (AL) distribution and to carry out the estimation via maximum-likelihood approach. The AL distribution can be written as

$$f(y_i|\mathbf{x}_i, \sigma, \tau) = \exp\left[-\rho_\tau\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}(\tau)}{\sigma(\tau)}\right)\right]\frac{\tau(1-\tau)}{\sigma(\tau)} \tag{2.8}$$

It can be shown that maximizing the log-likelihood of this distribution function corresponds to minimizing the objective quantile function defined in (2.7) (Koenker and Machado, 1999). Moreover, the loss function ρ_τ of (2.8) assigns weights τ or $(1-\tau)$ to the observations greater or less than $\mathbf{x}_i^T\boldsymbol{\beta}(\tau)$, respectively. The distribution is therefore split along the scale parameter into two parts, one with probability τ to the left, and one with probability $(1-\tau)$ to the right. These features have made the AL distribution suitable to be applied to quantile inference (Koenker and Machado, 1999; Yu and Moyeed, 2001; Yu et al., 2003; Geraci and Bottai, 2007; Liu and Bottai, 2009; Yue and Rue, 2011).

[10] Estimating the mean can be viewed as a problem of minimization of the sum of squared residuals. The squared distance function is differentiable at all points.

2.4.3 Properties and features

Different advantages make quantile regression appealing and are contributing to its growing popularity.

Equivariance

In addition to the classical equivariance properties, quantiles enjoy a stronger property, called *equivariance to monotonic transformations*. Let h be a non-decreasing function, then for any Y,

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau)) \tag{2.9}$$

In words, the quantiles of the transformed random variable $h(Y)$ are the transformed quantiles of the original Y. For example, a conditional quantile of $\log(Y)$ is the log of the conditional quantile of Y:

$$Q_{\log(Y)}(\tau) = \log(Q_Y(\tau)) \tag{2.10}$$

This property is of particular importance in the presence of skewed distribution where a logarithmic transformation is often used.

Robustness

Because of the nature of the minimized distance function, quantile regression estimates are not sensitive to outliers, that is, even a dramatic change in the highest measurement of a distribution would not influence the median. On the contrary, the mean of a distribution is sensible to even small changes in the outliers of the distribution.

Inference

In quantile regression approaches the bootstrap procedure (Efron and Tibshirani, 1994) is commonly preferred to derive standard errors for the estimates, given that the assumptions for the asymptotic procedure do not always hold. The bootstrap procedure offers strong flexibility to obtain standard errors and confidence intervals for any estimates or combination of estimates.

Softwares

Quantile regression can be performed with any major statistical software. Stata provide four commands to run quantile regression (`qreg`), interquantile range regression

(`iqreg`), simultaneous-quantile regression (`sqreg`), and bootstrap quantile regression (`bsqreg`). The same name (*quantreg*) is then use for both a SAS procedure, and a R library.

Result presentation

It is important to mention that the possibility of investigating the entire distribution of the outcome, which is one of the main advantages of the approach, could introduce some challenges when it comes to results presentation. When quantile methods are used we do not have a single summary measure to describe the entire association of interest, and a graphical presentation of the results is recommended as the optimal tool to summarize research findings, plotting results as a function of the corresponding quantile.

2.5 Quantile regression for censored data

2.5.1 Overview

After its first development in the past century, the quantile regression framework has been largely expanded and established methodologies are available for bounded outcomes (Bottai et al., 2010) and longitudinal studies (Geraci and Bottai, 2014). Various methods have been also introduced to deal with survival outcomes.

Time-to-event data are complicated by censoring, as we are usually interested in a response variable T_i but we only happen to observe $y_i = \min(t_i, c_i)$. In this setting, to estimate the parameters of a quantile regeression model (2.6) we need to minimize the following sum of residuals, introduced in Powell (1986):

$$\sum_{i=1}^{n} \rho_\tau(y_i - \min\{\mathbf{x}_i^T \boldsymbol{\beta}, c_i\}) \tag{2.11}$$

One important methods to fit censored quantile regression models was developed by Portnoy (2003), and is an estimation procedure that can be seen as a generalization of Kaplan-Meier to conditional quantiles. This method, however, makes the strong assumption that the regression models at lower quantiles are all linear (*global linearity*). Peng and Huang (2008) presented an alternative to this model based on the Nelson-Aalen estimator of the cumulative hazard function, which makes the same assumption of global linearity. These two methods were compared to a previous approach (Powell, 1986) and showed similar performances and a modest efficiency (Koenker, 2008).

Another valuable method was introduced by Wang and Wang (2009), who proposed

using the local Kaplan-Meier method to overcome the global-linearity assumption of the Portnoy's method. This approach is simpler but requires additional assumptions that could affect the estimates.[11] Other methods that have been proposed only apply to limited settings (Koenker and Geling, 2001; Bang and Tsiatis, 2002).

In 2010 Bottai and Zhang proposed to take advantages of the important properties that the AL distribution provides to quantiles estimation, and introduced Laplace regression for censored data (Bottai and Zhang, 2010).

2.5.2 Laplace regression

The aim of Laplace regression is to estimate the conditional quantiles of a time variable T_i. When T_i is censored we observe $y_i = \min(t_i, c_i)$, and $\delta_i = I(t_i \leq c_i)$.

The goal is to estimate the τth conditional quantile of T_i. Laplace regression assumes that

$$t_i = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) + u_i \tag{2.12}$$

where u_i follows the AL distribution introduced in (2.8) and $\mathbf{x}_i$ is a vector of covariates.

Estimation of the model parameters is conducted via maximum likelihood. The originally proposed algorithm for estimating the model has been recently improved after the introduction of a gradient search algorithm of maximization (Bottai et al., 2015). Mathematical details on the Laplace model and estimation are reported in Appendix 1.

2.5.3 Properties and features

Laplace regression assumes the errors to follow an AL distribution because of the discussed advantages of this distribution in the estimation of quantiles. Nevertheless, this parametric assumption, shared by other methods in quantile regression (Liu and Bottai, 2009; Farcomeni, 2010; Lee and Neocleous, 2010; Yuan and Yin, 2010), has been shown not to influence the performances of the model under different data distributions (Bottai and Zhang, 2010; Bottai and Orsini, 2013). Simulations studies have documented good performances of the model in terms of computational speed, precision, robustness of standard errors, and coverage of confidence intervals that was close to the nominal value (Bottai and Zhang, 2010; Bottai et al., 2015).

[11] The Wang & Wang approach requires the subjective choice of a smoothing function and performs better if independency between covariates can be assumed.

Compared to the other available methods for censored quantiles, these have a wider availability, being implemented in the official packages of R and SAS. However, Laplace regression provides considerable advantages in terms of computational speed and stability, performances that have been observed in different conditions and regardless of the assumed distribution (Bottai and Orsini, 2013; Bottai et al., 2015). Moreover, the other available methods for censored quantile regression do not make distributional assumption but are often subject to equally strong assumptions, such as the one of global linearity (Portnoy, 2003; Peng and Huang, 2008), not required by Laplace. Laplace regression is available as a user-friendly Stata command (Bottai and Orsini, 2013). The introduction of this Stata procedure was crucial to allow the possibility of applying the method on large dataset, as the implemented procedure to derive robust standard errors with the asymptotic theory[12] is considerably faster than the bootstrap. Appendix 2 provides a tutorial on the use of this Stata command.

It is also interesting to point out that in a sample where all observations are uncensored, the estimation equations are identical to those of the traditional quantile regression, as demonstrated in section 2.3 of Bottai and Zhang (2010).

Estimation of a single quantile of the distribution of T_i does not depend on the estimation of other quantiles, despite weighted information from the entire sample are involved in the estimation. Nevertheless, it is possible to simultaneously estimate Laplace regression models at different quantiles, and to provide statistical tests for differences between coefficients within and between different quantiles.

2.5.4 Laplace regression in epidemiology: modeling survival percentiles

The quantiles of a time-to-event variable T correspond to the survival percentiles defined in Section 2.3. This implies that statistical methods for censored quantiles, such as Laplace, can be used to statistically model survival percentiles as a function of covariates.

A common research goal in prospective cohort studies is to evaluate the effect of an exposure E on the time until developing an adverse event D. Laplace regression can be used to directly model the pth survival percentile of the time variable T as a function of E and other possible predictors (Orsini et al., 2012). To simplify the mathematical notation we can write a Laplace regression model on the pth survival percentile of T, conditional on E, as

[12] Asymptotic standard errors were not described in the original work from Bottai & Zhang and were introduced in Bottai and Orsini (2013).

$$T(p|E = e) = \beta_{p0} + \beta_{p1} \cdot e \tag{2.13}$$

In the simple situation where E is a dichotomous predictor, $\hat{\beta}_{p0}$ estimates the time by which $p\%$ of participants with $E = 0$ experience the event (t_0 from figure 2.5), while the time by which the same fraction of events is attained by participants with $E = 1$ is estimated by $\hat{\beta}_{p0}+\hat{\beta}_{p1}$. It naturally follows that $\hat{\beta}_{p1}$ is an estimate of the pth PD$=t_1-t_0$, and indicates the difference in time by which participants with $E = 1$ and $E = 0$ experience the same fraction of events. In this simple situation (one dichotomous exposure with no additional covariate included), survival percentiles estimated with Laplace correspond to those calculated with Kaplan-Meier. When E is continuous and is included in the model assuming a linear relationship with the pth survival percentile, the PD estimated by $\hat{\beta}_{p1}$ is interpreted as the difference in time (i.e. years, months, days) by which the specific proportion of events is achieved, for any one-unit increase in E. The linearity assumption in the relationship between a continuous predictor and survival percentiles can be relaxed by modeling exposures with flexible transformations such as polynomials or splines.

The model can be further expanded by including additional covariates, thus estimating multivariable-adjusted PDs. Multiple percentiles can be estimated simultaneously, testing coefficients within and between survival percentiles. The simultaneous estimation and plotting of different percentiles, similar to the one reported in Figure 2.6, might require some smoothing, depending on the variability across percentiles. In that case algorithms such as the *lowess* can be applied. A recent approach proposed by Frumento and Bottai (2015) allows for modeling of the quantile regression coefficients.

3. Aims of the thesis

The overall aim of this thesis was to introduce the percentile approach to time-to-event outcomes in epidemiology, by evaluating the impact of lifestyle factors on survival in a large prospective cohort of Swedish men and women, and by investigating and further developing the statistical framework for the estimation of conditional survival percentiles.

Specific aims of the epidemiological studies on lifestyle and overall mortality were:

- To assess the dose-response relationship between FV consumption and mortality from all-causes presenting the findings in terms of differences in survival percentiles (Study I).
- To evaluate the association between sleep duration and mortality across categories of total physical activity presenting survival differences associated with combined levels of sleeping and total physical activity (Study II).
- To investigate differences in survival across levels of total red meat consumption, and combined levels of processed and non-processed red meat intake (Study III).

Specific aims of the methodological studies were:

- To present the consequences of changing time-scale from follow-up time to attained age on the interpretation and estimation of the survival function, and examine the use of different time scales in the estimation of survival percentiles (Study IV).
- To investigate the insights that evaluating survival percentiles provides in the important epidemiological topic of interaction analysis (Study V).

4. Materials and methods

4.1 Study population

Study I, Study II, and Study III, whose specific aim was to assess the impact of lifestyle factors on survival, used data from a large population-based cohort of Swedish men and women. The study population was obtained by combining participants in the Cohort of Swedish Men (COSM), and the Swedish Mammography Cohort (SMC). This same population was also used in Study IV and Study V in the empirical example sections.

A detailed presentation of COSM and SMC is described in Harris et al. (2013). In brief, SMC included all women born between 1914 and 1948 and residing in two counties of central Sweden (Västmanland and Uppsala), who between 1987 and 1990 received a questionnaire and were invited to participate in the study. The questionnaire included questions on anthropometric measures, (e.g. body weight and height), socio-demographic information (e.g. educational level, marital status), woman's health status, and an extensive section with question related to lifestyle and diet. In the late fall of 1997, a second questionnaire was sent to participants who were still alive and residing in the study area. The aim of this second questionnaire was to update all information recorded with the first questionnaire, and to collect additional information such as smoking status, physical activity, and other lifestyle factors. The questionnaire was self-reported, and each participant was asked to report diet and lifestyle information over the previous year. The dietary section consisted of a detailed 96-item food frequency questionnaire. In total, 39,227 women (70%), aged 48 to 83, returned this second questionnaire.

The COSM was established in 1997 when all men residing in two regions of Central Sweden (Västmanland and Örebro), between the age of 45 and 79 years, were invited to respond a questionnaire and invited to participate in the study. To facilitate the combination and joint analysis of the two study populations, the self-administrated questionnaire received by men was specular to the one that women in the SMC received in 1997, with exception of items related to women's health. A total of 48,850 (49%) men returned the questionnaire.

The total size of the population investigated in these studies included the 88,077 men and women who participated in the 1997 SMC and COSM. Each study was subject to specific inclusion criteria, and for Study I, Study II, and Study III, the included sample sizes consisted of, respectively, 71,706 (38,221 men and 33,485 women), 70,973 (37,846 men and 33,127 women), and 74,645 participants (40,089 men and 34,556 women).

All variables recorded from the questionnaire underwent internal validation studies, which were conducted by assessing detailed weekly records in a sub-sample of the original population. The Spearman correlation coefficients between the average of four 1-week diet records and the dietary questionnaire ranged from 0.4 to 0.7 for individual fruit and vegetable items, and for items related to the consumption of red meat (Wolk, Unpublished data). Information on physical activity and sleeping were validated using two 7-day activity records that were performed 6 months apart in different subgroups, and showed a good correlation with the questionnaire variables (Spearman's rank correlation 0.6) (Orsini et al., 2008; Norman et al., 2001).

Both COSM and SMC were approved by the Regional Research Ethics Board at Karolinska Institutet, and all participants gave their informed consent.

4.2 Outcome assessment

The outcome examined in all the studies included in this thesis was mortality for all causes. Information on death was ascertained by linking COSM and SMC to the Swedish Register of Death Causes at the National Board of Health and Welfare. The link with the registries substantially increases the quality of the cohorts, as it has been documented that 93% of all deaths in Sweden are reported within 10 days, while 100% are reported within 30 days (Ludvigsson et al., 2009).

For Study I, at the time of performing the statistical analysis, mortality data from the register were available up to December 2010. The follow-up for this study was therefore closed at 31 December 2010, including a total of 13 years (January 1998 - December 2010). During this period 11,439 deaths occurred (6,803 men and 4,636 women). At the time of conducting the analyses for Study II and Study III two additional years of registry data were available, and follow-up was extended to the end of 2012. A total of 14,575 deaths (8,436 men and 6,139 women) and 16,683 deaths (6,948 women and 9,735 men), were respectively reported over the 15 years of follow-up.

4.3 Exposures assessment

The main exposures investigated in Study I, Study II, and Study III, were, respectively, fruit and vegetables (FV) intake, the combined effect of physical activity (PA) and sleeping, and processed and unprocessed red meat consumption. All these exposures were assessed using information from the the self-reported questionnaire.

Study I

Information about the daily intake of FV in the study population was obtained by using 14 questions on vegetables consumption (carrot, beetroot, lettuce, cabbage, cauliflower, broccoli, tomato, pepper, spinach, peas, onion, garlic, pea soup, other vegetables), 5 on fruits (orange, apple, banana, berry, other fruits) and one on orange juice. Total consumption of FV was summarized into a single continuous variable, expressed as servings/day, which was obtained by converting the questionnaire responses to an average daily intake of each item, and by adding the intake of all items together. A total of 46% of the participants completed all 20 questions on FV, and almost 80% reported less than 2 missing values. When aggregating items, it was assumed that missing values for an individual food meant no intake for that particular item (Hansson and Galanti, 2000).

Study II

Sleep duration was assessed with a single item through which participants were asked to report the average duration of sleep per day. This information was then coded into 5 categories (<6 hours per day, 6-6.5 hours per day, 7 hours per day, 7.5-8 hours per day, >8 hours per day).[1] Information on PA was obtained by asking the amount of five different domains of PA. These items were composed of six predefined activity levels for the item of work/occupational activity (from mostly sedentary to heavy manual labor) and five to six predefined categories for the other four domains: home/household work (from less than 1 hour to more than 8 hours per day), walking/bicycling (from hardly ever to more than 1.5 hours per day), inactive leisure-time (i.e. watching TV/reading, from less than 1 hours per day to 6 hours per day or more), and exercising (active leisure-time, from less than 1 hour to more than 5 hours per week). Based on the compendium of physical activities, each domain was assigned a specific intensity defined as metabolic equivalents (MET, kcal/kg - Norman et al. (2001)). The score calculated for each activity was then multiplied for its reported duration (hours) and a total daily score of PA was obtained by adding all specific activities together.

[1]Information on sleep duration was collected as a continuous covariate, but eventually categorized due to the high number of responses at specific values of the exposure.

Study III

The total intake of red meat, assessed in grams per day, was calculated by combining information on amount and frequency in the consumption of different types of red meat. Participants had to report the average frequency of consumption of different types of processed and non-processed red meat, by using 8 predefined frequency categories ranging from never to 3 or more times per day. Non-processed red meat included fresh and minced pork, beef, and veal. Processed red meat included sausages, hot dog, salami, ham, processed meat cuts, liver pate, and blood sausage. Together with the frequency, participants were asked to report, for each item, the average portion size. All reported information were combined to derive continuous variables of total, processed, and unprocessed meat intake. The age of the participants was taken into account when translating frequency and portion sizes into the total information of red meat intake.

4.4 Exposures modelling

The continuous exposures evaluated in Study I and Study III were flexibly modeled by means of cubic splines. When dealing with continuous exposures, a simple inclusion of a covariate in the statistical model would assume that the dose-response association between the exposure and the outcome is linear. A common option to relax this assumption is to categorize the continuous covariate, but this approach presents different limitations and often rely on unrealistic assumptions (Greenland, 1995a,c; Royston et al., 2006). Categorization of quantitative predictors, by pooling participants into groups, implicitly assumes a *step function* shape for the dose-response, with a variation in the risk only at specific values of the exposure. A consequence of this assumption is the loss of within-category information, as it is not possible to detect differences between individuals grouped in the same category. A second critical drawback of categorization is the subjective choice of cutoff points, which implies a loss of statistical power and may dramatically influence the results.[2]

Spline transformations are a common tool to model a continuous covariate relaxing the assumption of linearity, though avoiding the loss of information implied by the use of categories (Greenland, 1995b; Royston et al., 1999; Steenland and Deddens, 2004; Marrie et al., 2009). The basic idea behind splines is to split the exposure distribution into a predefined number of intervals, placing the splits at points called *knots* (Durrleman and Simon, 1989). The most common choice is to model each interval of the

[2]The following website provides excellent interactive graphs to visually explain the limitations of the categorical approach and the advantages of using splines: http://www.le.ac.uk/hs/pl4/spline_eg.html

curve with cubic polynomials (*cubic splines*). Formally, a cubic spline transformation for a predictors x, modeled with n knots values $k_i, i = 1 \dots n$, takes the form:

$$G(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{i=1}^{n} \beta_{3+i} \max(x - k_i, 0)^3 \tag{4.1}$$

This function is constrained to join at the knots locations (continuity of first derivative), and continuity of the second derivative can also be assumed to improve smoothness. In addition it can be forced to be linear before the first (*left-restricted*) or after the last knot (*right-restricted*). For example, a spline transformation with a linear restriction on the right tail takes the form:

$$G(x) = \beta_0 + \beta_1(-x) + \sum_{i=1}^{n} \beta_{1+i} \cdot \max(k_i - x, 0)^3 \tag{4.2}$$

When both the right and the left tail are constrained to linearity (*restricted cubic splines*) the function with 3 knots is simplified to two regression transformations:

$$G(x) = \beta_0 + \beta_1 \cdot G_1(x) + \beta_2 \cdot G_2(x) \tag{4.3}$$

where

$$G_1(x) = x$$

and

$$G_2(x) = \frac{(x - k_1)_+^3 - (x - k_2)_+^3 \cdot \frac{k_2 - k_1}{k_3 - k_2} + (x - k_3)_+^3 \cdot \frac{k_3 - k_1}{k_3 - k_2}}{(k_3 - k_1)^2}$$

Fitting a restricted cubic splines function can be easily accomplished by different statistical software and can be summarized as a two-step procedure (Orsini and Greenland, 2011). First, the transformations of the original covariate must be generated. Second, the new variables should be included in the model in the place of the original covariate. A simple tool to assess departures from linearity, in the restricted cubic splines setting, is to test whether the coefficient of the second spline is equal to zero. When right-restricted or left-restricted cubic splines are used, linearity can be evaluated by testing the null hypothesis that the coefficients of the unrestricted spline transformations are jointly equal to zero.

4.5 Statistical analysis

Study I

The association between FV consumption and mortality was modeled in terms of survival, and Laplace regression was used to calculate multivariable adjusted PDs across levels of FV. Over the 13 years of follow-up 16% of the study population experienced the event of interest (mortality for all-causes). To minimize data extrapolation and to ease results interpretation we therefore decided to focus the main analyses on the 10th survival percentile, presenting association in terms of 10th PDs. Other observed percentiles were evaluated as additional analyses and provided similar results.

Possible confounders of the association that were included in the multivariable model comprised gender, baseline age, body mass index (BMI), total physical activity, smoking status and pack-years of smoking, alcohol consumption, education level, and total energy intake.

To investigate the dose-response relation between FV consumption and survival, and to fully catch the effects at low levels of FV consumption, the main exposure of daily FV consumption was flexibly modeled using right-restricted cubic splines with three knots of the distribution (at 3, 5, and 8 servings of FV per day). Percentile differences were presented using the recommended dose of 5 servings per day as referent.

Analyses were performed in Stata, version 12 (StataCorp).

Study II

The proportion of cases reported during follow-up, after the inclusion of two additional years, was close to 20%. The main analyses of this study were therefore focused on the 15th percentile of survival. Other percentiles were evaluated in additional analyses and no substantial differences in the results were observed. Differences in survival were estimated by fitting multivariable-adjusted Laplace regression models on the 15th survival percentiles. All multivariable analyses were adjusted for sex, age at baseline, BMI, smoking status and pack-years of smoking, alcohol consumption, and educational level.

The overall association between categories of sleep duration (<6 hours per day, 6-6.5 hours per day, 7 hours per day, 7.5-8 hours per day, >8 hours per day) and survival was evaluated in a multivariable-adjusted model using the group of participants who slept 7 hours per day as reference category.

Next, the interaction between the categorical variables of sleep duration and physical activity (tertiles: <39.3, 39.3-44.2, >44.2 MET-hrs/day) was assessed. A p-value for interaction was obtained by simultaneously testing the product terms of the indica-

tor variables equal to zero. The association between sleep duration and mortality was then evaluated over tertiles of total PA.

All analyses were performed in Stata, Version 12 (StataCorp).

Study III

The proportion of cases observed during follow-up was similar to the one reported in Study II, and the main analyses were focused on the 15th percentile of survival. Evaluating other percentiles within the observed range did not influence the results.

First, the dose-response association between total red meat intake and survival was investigated. The main exposure was flexibly modeled by means of restricted cubic splines, with 3 knots at fixed percentiles of the distribution (31, 77, and 140.5 grams/day), and using no consumption of red meat as reference value.

To evaluate the joint association of processed and non-processed meat consumption in predicting survival, an interaction term between the two exposures was included in the model. The overall p-value for statistical interaction was obtained by testing the four regression coefficients of the interaction terms between splines (two cubic splines for processed meat multiplied by two cubic splines for non-processed meat) jointly equal to 0. For this analysis the median value of consumption for processed and non-processed red meat (30.6 grams/day and 44.5 grams/day, respectively) were used as reference group.

PDs across levels of processed and non-processed meat were estimated with Laplace regression models on the 15th survival percentile, further adjusting for baseline age, gender, BMI, total physical activity, smoking status, alcohol consumption, energy intake, educational level, fruit and vegetables consumption, and prevalence of diabetes.

All analyses were performed in Stata, Version 13 (StataCorp).

5. Results

5.1 Study I - Fruit and vegetables and survival

In this study the association between fruit and vegetables consumption was modeled in terms of survival time by using splines transformations. The overall 10th survival percentile was 116 months (95% CI: 114, 118), meaning that 90% of the cohort was still alive after 9.6 years of follow-up.

Figure 5.1 depicts the dose-response association between FV consumption and survival. A strong departure from linearity was detected, observing a substantial increase in survival up to 5-6 servings/day of FV. Higher levels of consumption were not statistically associated with an additional increase in survival. Compared with a FV consumption of 5 servings/day, lower levels were progressively associated with shorter survival up to 3 years for those who never consumed FV (10th PD= -37 months; 95% CI: -58, -16).

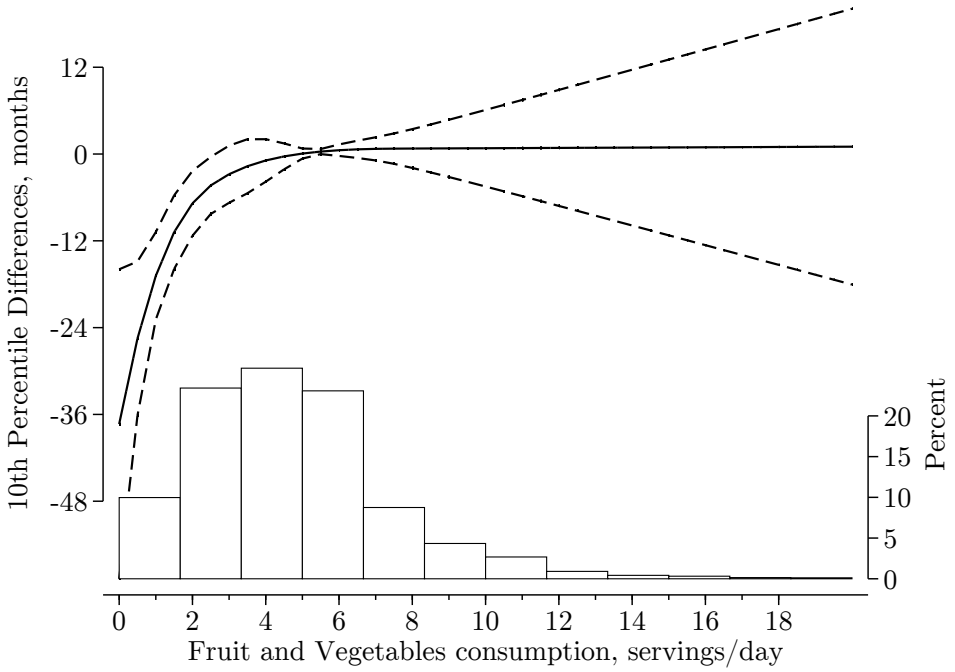


Figure 5.1: Multivariable adjusted 10th percentile differences (differences in months by which 10% of the cohort has died) as a function of fruit and vegetable consumption. Dashed lines represent 95% CIs. The reference value is 5 servings/dat, and the histogram shows the distribution of fruit and vegetable consumption in the cohort. The statistical model was adjusted for gender, baseline age, body mass index, total physical activity, smoking status and pack-years of smoking, alcohol consumption, education level, and total energy intake. The main exposure was modeled with right-restricted cubic splines.

5.2 Study II - Sleep duration and survival across levels of physical activity

This study aimed to assess the association between sleep duration and mortality across categories of PA. The overall association between sleep duration and mortality confirmed a pronounced U-shape, with participants sleeping an average of 7 hours per day experiencing the longest survival (Table 5.1). Compared to the reference group, 15% of men and women with the lowest (<6 hrs) and highest sleep duration (>8 hrs) died about 1 year earlier (15th PD= -12 months; 95% CI: -19, -6 and 15th PD= -20 months; 95% CI: -26, -14, respectively).

A significant interaction between sleep duration and PA in predicting mortality was observed (p for interaction <0.001), and the association between sleep duration and mortality was different across levels of PA (Table 5.1). Short sleep duration, less than 6 hours per day, was associated with shorter survival throughout the entire distribution of PA. Long sleep duration, more than 8 hours per day, was associated with significantly shorter survival (15th PD= -20 months; 95% CI: -30, -11) only among those in the lowest tertile of PA. No significant associations were observed between long sleep duration and mortality for those with medium or high levels of PA.

Table 5.1: 15th Percentile differences in survival according to categories of sleep duration, overall, and stratified by tertiles of total physical activity

Mean Sleep Duration, hours/day[a]	**N**	**Cases**	**15th PD**	**95% CI**
Overall				
<6	2,982	882	-12	(-19, -6)
6–6.5	13,194	2,812	-4	(-9, -2)
6.6–7.4	27,891	4,599	0	Referent
7.5–8	22,972	5,032	-3	(-9, -1)
>8	3,934	1,250	-20	(-26, -14)
First tertile[b]				
<6	316	109	-27	(-51, -3)
6–6.5	2,309	495	-10	(-18, -2)
6.6–7.4	7,375	1,147	0	Referent
7.5–8	7,167	1,525	-4	(-10, 2)
>8	1,563	549	-20	(-30, -11)
Second tertile[c]				
<6	708	182	-15	(-26, -4)
6–6.5	3,700	673	0	(-7, 7)
6.6–7.4	7,618	1,148	0	Referent
7.5–8	5,811	1,164	-5	(-11, 0)
>8	886	226	-7	(-21, 8)
Third tertile[d]				
<6	1,135	286	-11	(-21, -1)
6–6.5	4,375	839	-8	(-14, -2)
6.6–7.4	7,735	1,211	0	Referent
7.5–8	7,167	967	-3	(-9, 3)
>8	484	107	-5	(-19, 9)

[a]Median values for sleep duration are 5, 6, 7, 8, and 9 hours/day for the categories of less than 6, 6 6.5, 6.6–7.4, 7.5–8, and more than 8 hours/day, respectively.

[b]First tertile (low physical activity) = <39.3 (median, 37) MET hours/day.

[c]Second tertile (medium physical activity) = 39.3–44.2 (median, 42) MET hours/day.

[d]Third tertile (high physical activity) = >44.2 (median, 47) MET hours/day.

5.3 Study III - Processed and unprocessed red meat consumption in predicting mortality

This study's aim was to investigate the combined role of processed and unproccesed red meat in predicting mortality. The dose-response association between red meat and survival showed a significant departure from linearity (p-value<0.001)(Figure 5.2). A considerable decrease in survival was only observed at levels of consumption higher than 100 g/day of red meat. Compared with participants who never consumed red meat, those consuming 200 grams/day lived about 1 year shorter (15th PD= -10 months; 95% CI: -18, -3). Increased consumption up to 300 g/day was associated with almost 2 years of shorter survival (15th PD= -21 months; 95% CI: -31, -10).

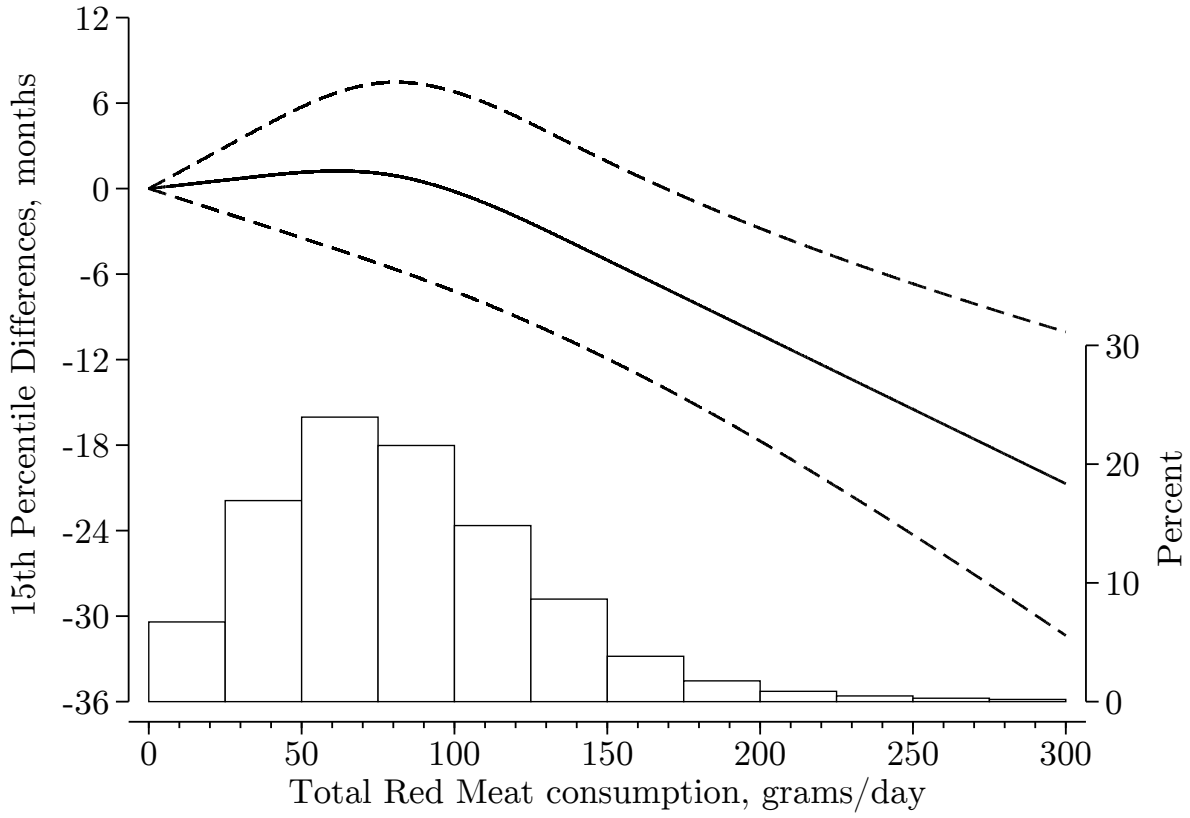


Figure 5.2: Multivariable adjusted 15th percentile differences (differences in months by which 15% of the cohort has died) as a function of red meat consumption, with the null consumption as referent. Dashed lines represent 95% CIs. The histogram shows the distribution of red meat consumption in the cohort. The statistical model was adjusted for baseline age, gender, BMI, total physical activity, smoking status, alcohol consumption, energy intake, educational level, fruit and vegetables consumption, and prevalence of diabetes. The two exposures were modeled by means of restricted cubic splines.

When investigating the joint effect of processed and non-processed red meat in predicting survival we observed a non-significant interaction between the two predictors (p-value for interaction=0.6). However, this did not prevent the association between non-processed red meat consumption and survival to be substantially different across levels of processed red meat intake (Figure 5.3). Comparing to participants with median consumption of processed and non-processed meat, higher intake of non-processed meat was associated with shorter survival only if combined with higher processed meat consumption (when consuming 100 g/day of both processed and non-processed meat: 15th PD= -19 months; 95% CI: -37, -2). High processed meat consumption was instead associated with shorter survival regardless the consumption of non-processed meat.

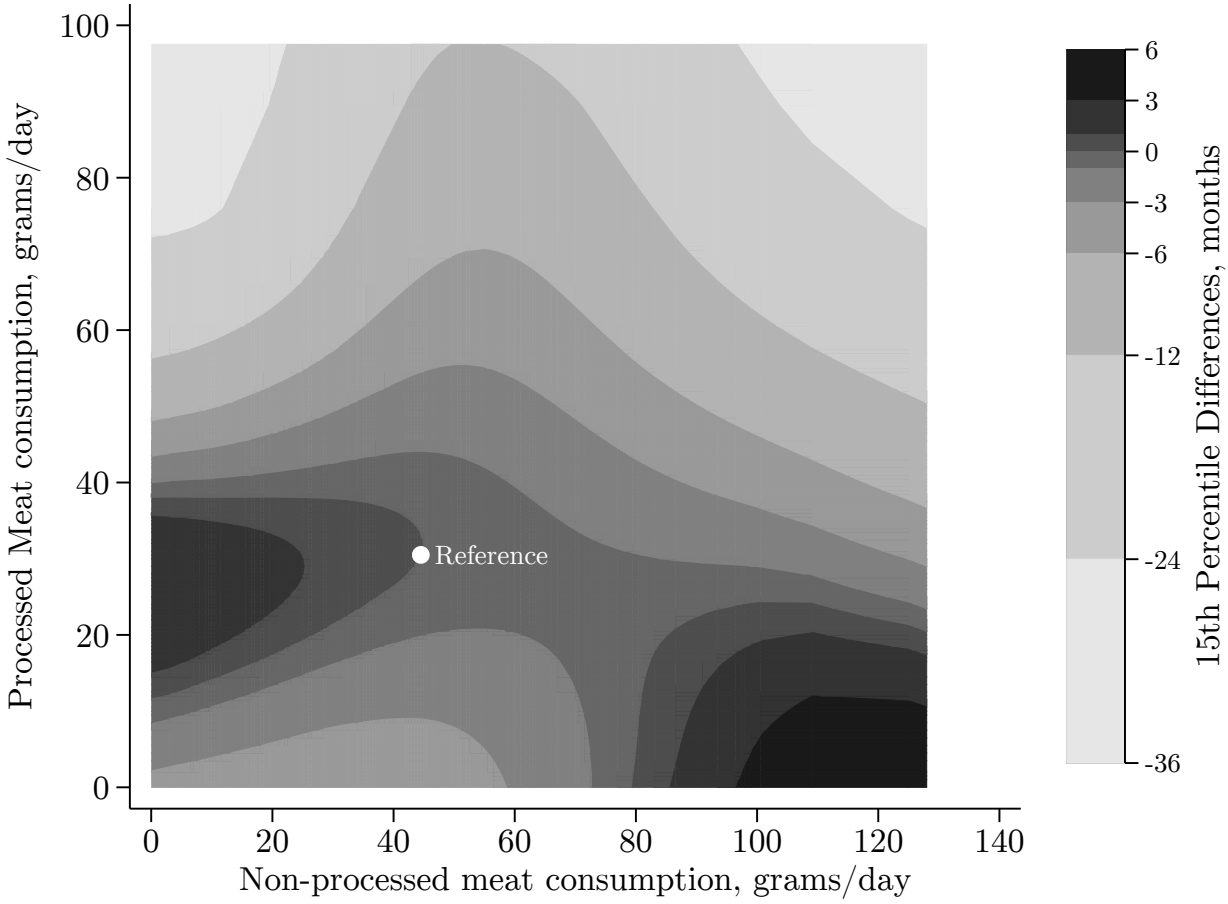


Figure 5.3: Multivariable adjusted 15th percentile differences as a function of combined levels of processed and non-processed red meat consumption. An interaction term between processed and non-processed meat is included in the model. The median consumption is used as reference. The model was adjusted for baseline age, gender, BMI, total physical activity, smoking status, alcohol consumption, energy intake, educational level, fruit and vegetables consumption, and prevalence of diabetes. The two exposures were modeled by means of restricted cubic splines.

5.4 Study IV

5.4.1 Consequences of changing the time scale

The description of the survival function provided in section 2.2, and the interpretation and estimation of survival percentiles as presented in section 2.3 and 2.5, are only valid if follow-up time is chosen as time-scale of the study. Changing the time scale to attained age strongly influences the way the survival curve is interpreted and estimated (Royston and Lambert, 2011; Lawless, 2011).

When follow-up time is used, all participants enter the study at the same time point and are followed until they experience the event of interest or are censored (top-left panel of Figure 5.4). The first consequence of changing time scale to attained age is that participants enter the study at different values of the time variable. Entries into the study are spread over the age range (*delayed entries*) and *left-truncation* is introduced.[1]

The presence of delayed entries complicates the interpretation of the survival curve, and changes the location of censored observations. Delayed entries are depicted in the top-right panel of Figure 5.4. The consequences of changing time-scale on the censoring mechanism are illustrated in the middle panels of Figure 5.4, which show a simple situation of a closed cohort observed for 15 years, during which 20% of the participants die, and the other 80% are regarded as censored observation. When focusing on follow-up time (middle-left panel), cases are distributed across the follow-up period, and all censoring occurs at the end, assuming no losses to follow-up. Any attempt to estimate survival percentiles above the 20th (such as the median survival) would require data extrapolation beyond the range of observed data. By changing the time scale to attained age (middle-right panel), censored observations distribute across the entire range of age, making possible the estimation of higher percentiles, up to the 99th if the oldest participant is a case.

Estimators of the survival function such as Kaplan-Meier are impacted by the change in the time-scale. When age is chosen as primary time scale the survival curve can still be estimated (bottom-right panel of Figure 5.4) but becomes hard to interpret, as it does not represent, at each age point, the age by which specific proportion of the populations have experienced the event of interest. Because of the presence of delayed entries, the Kaplan-Meier estimator and the survival function itself become conditioned

[1]Left-truncation defines a mechanism such that it is only possible to observe outcomes above the truncation limit. Differently from censoring mechanisms, truncation is due to a systematic selection process inherent to the study design.

on having survived up to the earliest truncation time (Mackenzie, 2012).

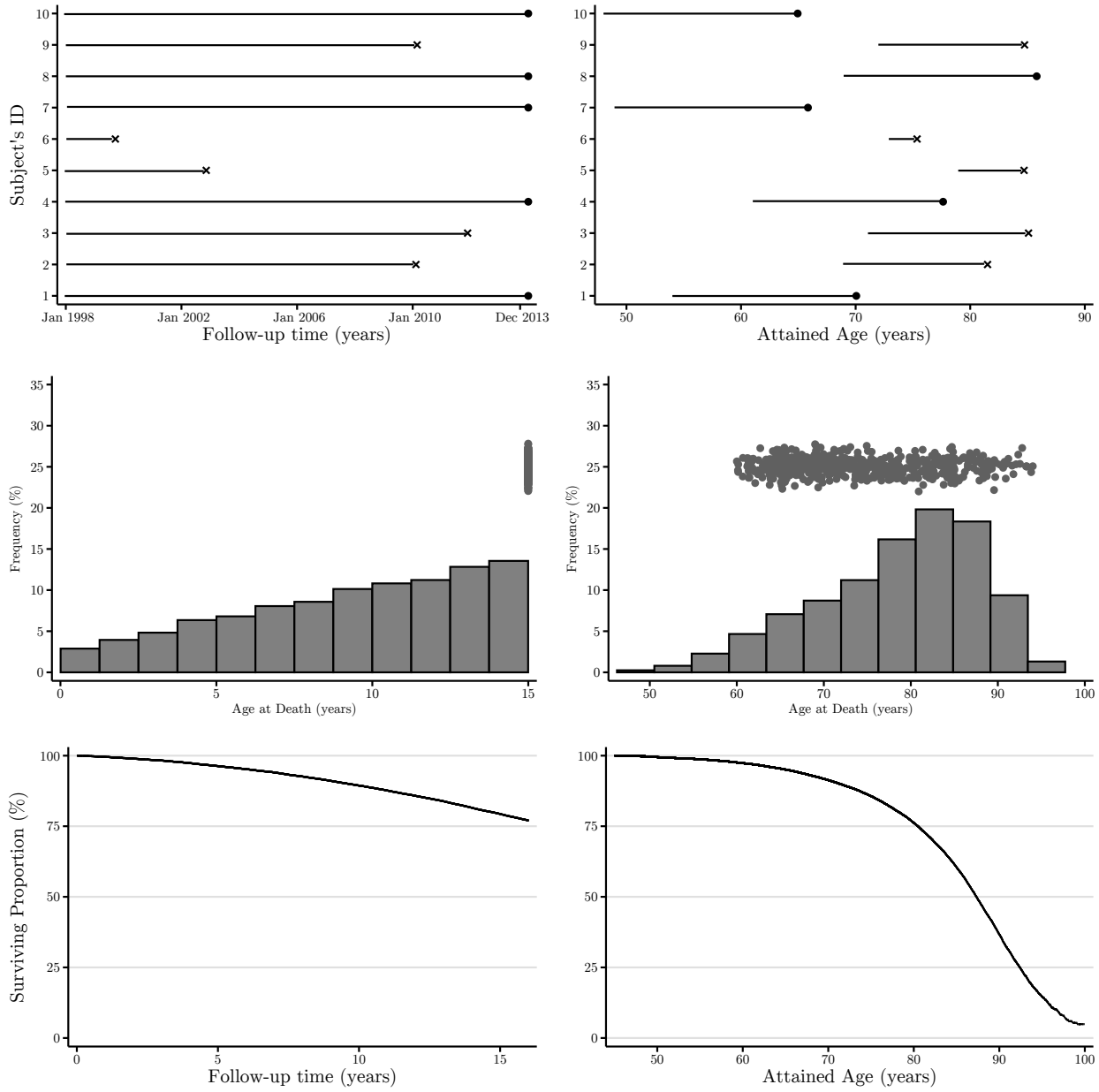


Figure 5.4: Consequences of changing the time scale from follow up time (left-column) to attained age (right column) on participants' entry (first row), cases and censoring distribution over time (second row; histogram, cases: gray dots, censoring) and survival curve (third row).

5.4.2 Interpretation and estimation of survival percentiles

Despite losing its interpretation, the curve depicted in the bottom-right panel of Figure 5.4 maintains the mathematical properties of a survival function as presented in Section 2.2. Therefore, the curve can be seen as a summary of the percentiles of the time variable of attained age at the event. The interpretation of these percentiles, despite conditioning on having survived up to the earliest truncation time, remains hard. The pth percentile of age is not interpreted as the age by which p% of the study population has experienced the event of interest. It may occur, for example, that by age a, less than p% of participants have actually entered the study.

An intuitive way of taking into account the presence of delayed entries, simplifying the interpretation of the percentiles of attained age, is to condition each individual survival experience on his/her baseline age. Let us assume a simple situation in which participants may enter the study at three different baseline ages (55, 65, or 75 years). Figure 5.5 depicts the survival curves for the three different sub-populations, defined by their baseline age, and assuming attained age at the event as time scale. In this situation, given a percentile p (in the figure p=25), the age points a_{1-3} can be interpreted as the ages by which p% (25%) of participants in each group have experienced the event.

Following the notation introduced in (2.13), to build a statistical model for conditional percentiles of attained age we change the time variable from follow-up time T, to attained age A. A is censored, and we observe, for each participant i, $z_i = \min(a_i, b_i)$ - that is - the smaller value between the attained age at the event, a_i, and the attained age at the end of follow-up (or at censoring), b_i.

By fitting a censored quantile regression model, such as Laplace, on a given percentile of the distribution of attained age with the only the intercept β_0, we provide a crude estimate of the percentile, such as the median age at event when p=50. This estimate, however, corresponds to the one obtained with the Kaplan-Meier estimator using age as the primary time scale, and shares the same limitation previously described.

However, it is possible to include age at baseline as a covariate in the statistical model, thus conditioning the estimates of attained age at the event on the individual age at baseline. For example, using the simple situation presented above with three possible values of baseline age, a regression model on the pth survival percentile, adjusted for age at baseline, could be used to predict the same estimates a_{1-3} represented in Figure 5.5.

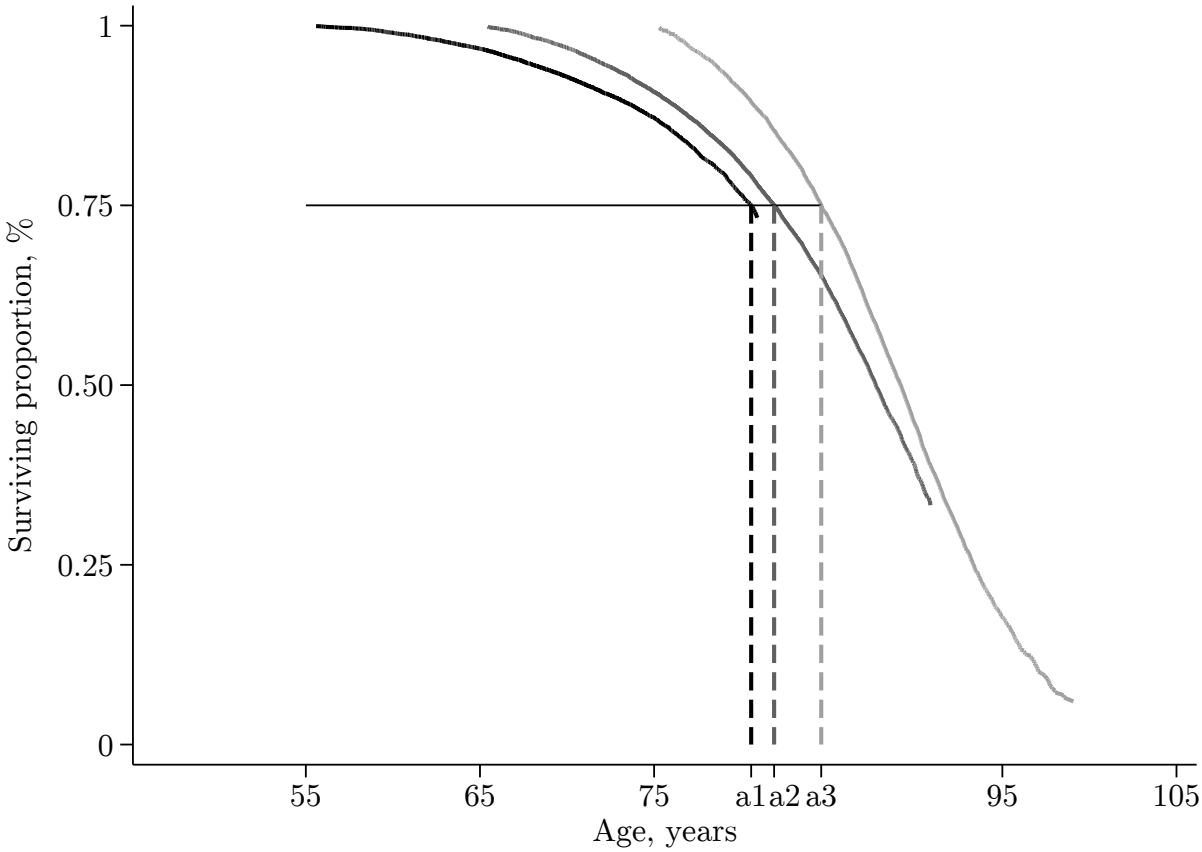


Figure 5.5: Survival curves, and survival percentiles, for three different groups of participants defined by their age at entry, when attained age at the event is chosen as primary time-scale.

Age at baseline is commonly assessed as a continuous covariate. It could be included in the model as a numerical predictor without any transformation, but this would assume a linear relationship between age at baseline and the percentiles of age at death. To relax this assumption, the inclusion of a function of age at baseline, such as a categorized covariate or a mathematical transformation, is commonly preferred. A Laplace regression model for the pth percentile of attained age takes the form:

$$A(p|\text{age\_baseline}) = \beta_{p0} + \beta_{p1} \cdot f(\text{age\_baseline}) \tag{5.1}$$

It is important to note that inclusion of a function of age at baseline implies that a certain fraction of extrapolation on some covariate patterns will be unavoidable, regardless the chosen percentile.

To assess the impact of an exposure E on the pth survival percentile, the model can be further conditioned on E:

$$A(p|E = e, \text{age\_baseline}) = \beta_{p0} + \beta_{p1} \cdot f(\text{age\_baseline}) + \beta_{p2} \cdot e \tag{5.2}$$

For a given p, the estimate of β_{p2} expresses the difference in the pth percentile of attained age between exposed and non-exposed participants, conditioned on baseline age. When $p = 50$, the model estimates differences in median age at the event. Further inclusion of other covariates would estimate multivariable-adjusted percentile differences. An implicit assumption of model (5.2) is that the effect of the main exposure on the pth percentile of age at death is constant across levels of age at baseline. This assumption might be relaxed by including in the model a term for interaction between baseline age and the exposure of interest.

5.4.3 Review of the results from Study III

To illustrate the use of Laplace regression to model percentiles of attained age at the event, the main result of Study III were replicated by evaluating differences in median age at death according to levels of total red meat consumption in the COSM/SMC cohort. Analyses from Study III were replicated step by step with the only difference that the outcome for each participant was defined as the age at death and that the Laplace regression model was fitted on the 50th percentile of this new time variable. While Figure 5.2 reported the main result by presenting the dose-response association between total red meat consumption and the 15th percentile of time since entry into the study, Figure 5.6 illustrates the dose-response association between total red meat consumption and the median age at death.

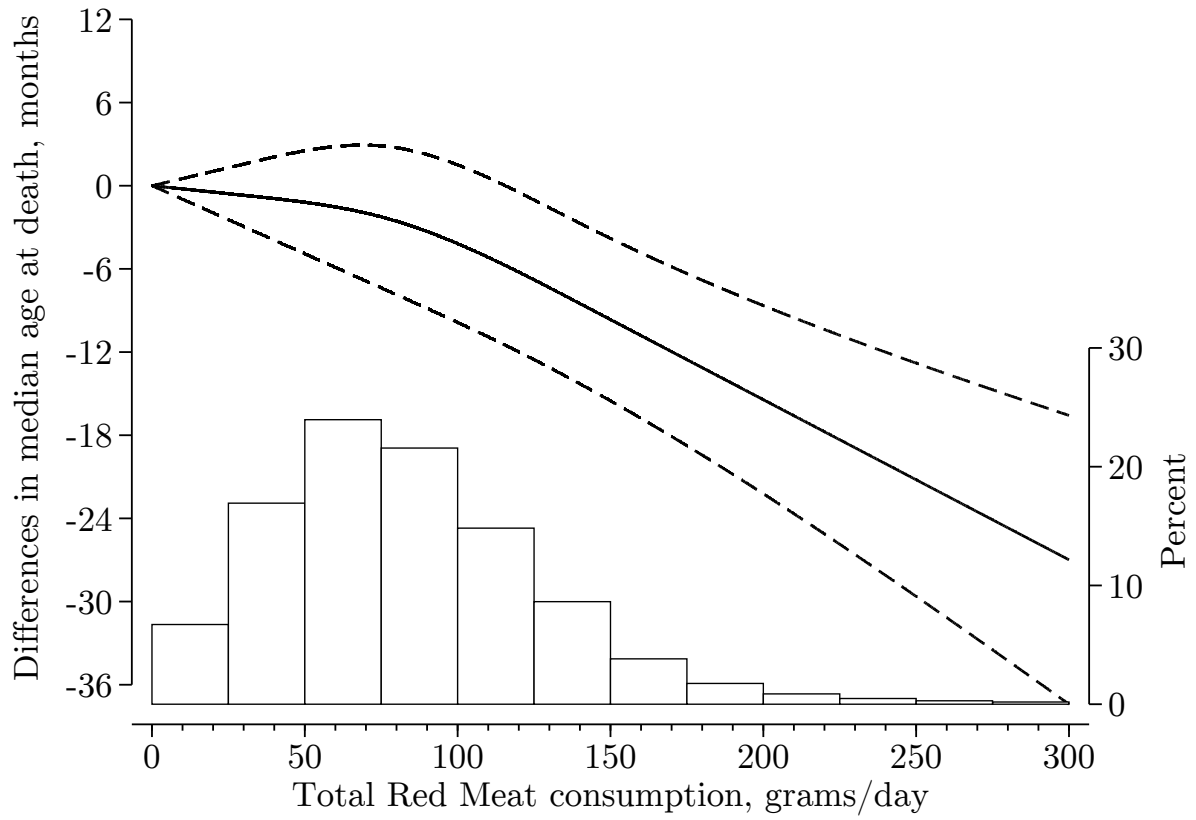


Figure 5.6: Multivariable adjusted differences in median age at death, with the null consumption as referent. Dashed lines represent 95% CIs. The histogram shows the distribution of red meat consumption in the cohort. The model was adjusted for baseline age, gender, BMI, total physical activity, smoking status, alcohol consumption, energy intake, educational level, fruit and vegetables consumption, and prevalence of diabetes. The two exposures were modeled by means of restricted cubic splines.

5.5 Study V

5.5.1 Interaction in epidemiology

Despite the rapid growth of epidemiological activity over the last decades, different epidemiological concepts are still underdeveloped and subject to methodological controversies (Rothman et al., 2008). One major example is the epidemiological concept of interaction, which is of primary interest to understand the main and combined effects of two concurrent risk factors in the development of a disease. The concept of interaction in epidemiology, and possible statistical tools to assess it in observational studies, have been investigated for over 40 years (Rothman, 1974). Nevertheless, interaction remains a popular topic in the epidemiological literature, and no agreement has been reached on its definition, neither established tools for its estimation are defined. Recently, the current literature on interaction analysis was reviewed in a textbook (VanderWeele, 2015), and summarized in a tutorial that currently represents the best available introduction to the topic (VanderWeele and Knol, 2014).

This final study aimed to investigate the advantages that evaluating survival percentiles provides to the topic of interaction analysis, by defining the concept of interaction between two exposures in the metric of time, and investigating its properties, estimation tools, meaning, and public-health implications.

5.5.2 Interaction assessment in time-to-event analysis

Interaction refers to the situation in which the effect of an exposure on the outcome may depend on the value of another exposure. It is commonly assessed as a departure from *additivity* or *multiplicativity* of the effects. When the combined effect of two exposures is the sum of the two main effects, than we are in the presence of additivity of the effects and there is no interaction on the *additive scale*. When the combined effect is equal to the product of the two main effects, we are instead observing a perfect multiplicativity of the effects, and we are in the absence of interaction on the *multiplicative scale*. It simply follows that, given these definitions, absence of interaction on one scale is likely to imply the presence of interaction on the other scale (Rothman et al., 1980; Greenland, 2009). Various studies have underlined the important public health meaning of additive interaction, which can be used to assess which subgroups of individuals are to be treated (Saracci, 1980; Rothman et al., 2008; Greenland, 2009; Knol et al., 2011; VanderWeele and Knol, 2014). In general, presenting both additive and multiplicative interaction would provide a complete picture of how two exposures

interact in predicting the outcome, and this procedure has been widely recommended (Botto and Khoury, 2001; Von Elm et al., 2007; Knol and VanderWeele, 2012; VanderWeele and Knol, 2014). However, despite these recommendations, this practice remains uncommon (Knol et al., 2009).

An important distinction must be made between biological and statistical interaction (Rothman, 1974; Rothman et al., 1980; Thompson, 1991; Greenland, 1993; Rothman, 1995). Statistical interaction, which is usually assessed by including in the regression model a product-term between the two exposures of interest (Greenland, 1983), arises from a statistical model and should not be used to draw biological conclusions (Siemiatycki and Thomas, 1981; Thompson, 1991; Cordell, 2002; Rothman et al., 2008). The evaluation and interpretation of interaction in epidemiological studies is strongly dependent on the scale of the model chosen to analyze data, which can be either additive or multiplicative (Rothman et al., 2008). Inclusion of a product term between two exposures in an additive model will serve as a test for additive interaction, while inclusion of a product term in a multiplicative model will test for departures to multiplicativity of the effects.

In time-to-event analysis, given the multiplicative nature of the Cox PH regression, interaction analysis is generally limited to the multiplicative scale. Measures to estimate additive interaction in time-to-event analysis have been proposed, the principal one being the relative risk due to interaction (RERI) (Li and Chambless, 2007), which can be calculated, after fitting a Cox model, with the following formula:

$$RERI_{HR} = e^{\beta_1+\beta_2+\beta_3} - e^{\beta_1} - e^{\beta_2} + 1 \tag{5.3}$$

The RERI, however, is subject to some limitations, as it is best suited in the presence of rare outcomes, and only provides indication of the direction of the interaction effect but not on its magnitude (Skrondal, 2003).

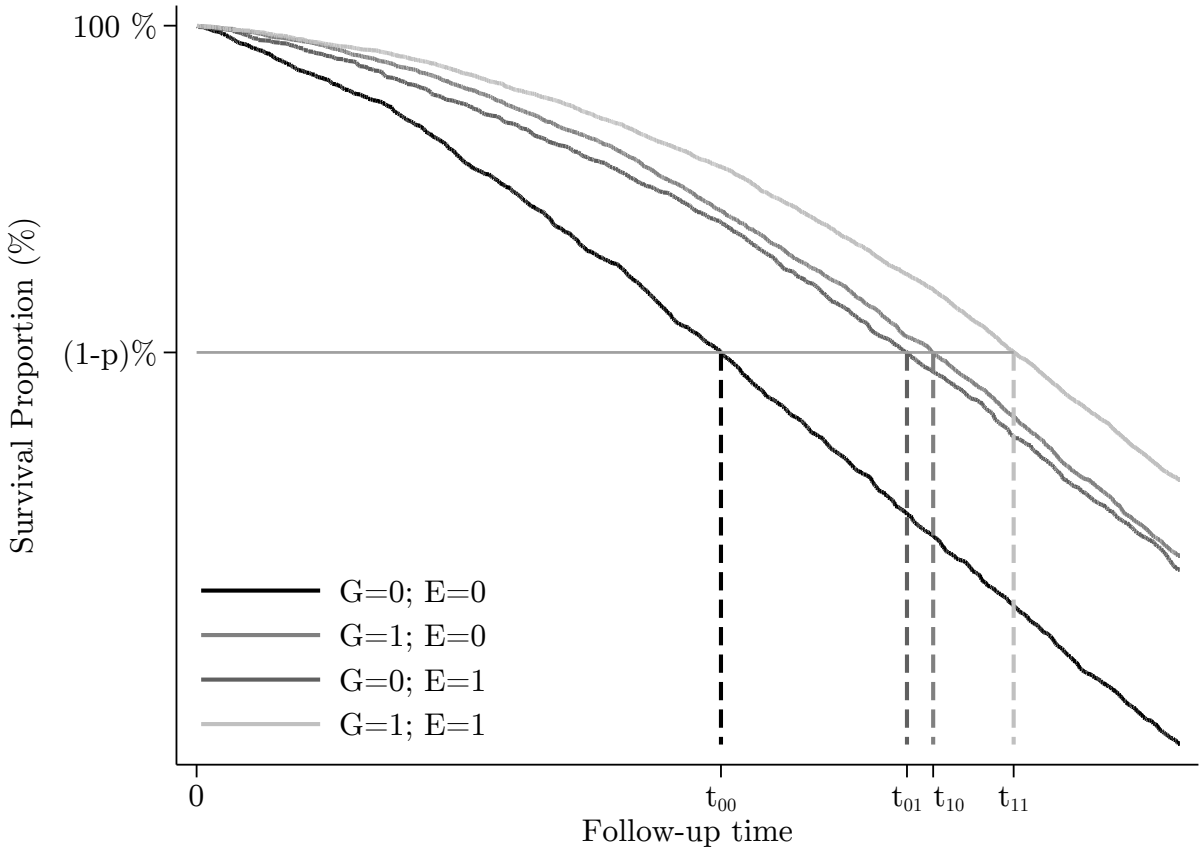


Figure 5.7: Survival percentiles according to combined groups of two dichotomous exposures E and G.

5.5.3 Additive interaction in the metric of time

Let G and E be two binary exposures, which can take values 0 or 1, and are both risk factors for the event D. Figure 5.7 presents a possible survival experience for the four combinations of the two exposures (i.e $G = 0, E = 0; G = 1, E = 0; G = 0, E = 1; G = 1, E = 1$). Given a fixed proportion of events p, the pth survival percentiles for each of the four groups $(t_{00}, t_{01}, t_{10}, t_{11})$ are displayed in the figure. The difference $(t_{11} - t_{00})$ represents the difference in the pth survival percentile between participants with both exposures and participants with none. The quantities $(t_{10} - t_{00})$ and $(t_{01} - t_{00})$ are PDs between participants with the only exposure G or E, respectively, and participants with none exposure. Following the conventional notation introduced in terms of risk (Rothman et al., 2008; VanderWeele and Knol, 2014), the following measure can be calculated to describe interaction between G and E at the pth percentile:

$$I_p = (t_{11} - t_{00}) - [(t_{10} - t_{00}) + (t_{01} - t_{00})] \tag{5.4}$$

This difference, calculated as an additive measure, can be rewritten as $t_{11} - t_{10} - t_{01} + t_{00}$ and represents a measure of additive interaction in the metric of survival time. It expresses to what extent the difference in survival due to the presence of both exposures exceeds the sum of survival differences due to each specific exposure. Comparing this measure with 0 we can define the interaction as *super-additive*, if greater than 0, or *sub-additive* if smaller than 0. This measure of interaction can be evaluated at any observed percentile, allowing to investigate how the interaction between two risk factors of interest is changing according to the fraction of events considered over time.

5.5.4 Model-based estimation

To evaluate the impact of the two binary exposures G and E and their interaction on the pth survival percentile of the time variable T in a regression-based framework, we can include a product term in a statistical model for the conditional pth survival percentile, such as Laplace regression:

$$T(p|G = g, E = e) = \beta_{p0} + \beta_{p1} \cdot g + \beta_{p2} \cdot e + \beta_{p3} \cdot g \cdot e \tag{5.5}$$

From (5.5) it is possible to estimate the pth survival percentiles for the four combinations of the two exposures, corresponding to the time points displayed in Figure 5.7.

$$\begin{aligned}
t_{00} = T(p|G = 0, E = 0) &= \beta_{p0} + \beta_{p1} \cdot 0 + \beta_{p2} \cdot 0 + \beta_{p3} \cdot 0 = \beta_{p0} \\
t_{10} = T(p|G = 1, E = 0) &= \beta_{p0} + \beta_{p1} \cdot 1 + \beta_{p2} \cdot 0 + \beta_{p3} \cdot 0 = \beta_{p0} + \beta_{p1} \\
t_{01} = T(p|G = 0, E = 1) &= \beta_{p0} + \beta_{p1} \cdot 0 + \beta_{p2} \cdot 1 + \beta_{p3} \cdot 0 = \beta_{p0} + \beta_{p2} \\
t_{11} = T(p|G = 1, E = 1) &= \beta_{p0} + \beta_{p1} \cdot 1 + \beta_{p2} \cdot 1 + \beta_{p3} \cdot 1 = \beta_{p0} + \beta_{p1} + \beta_{p2} + \beta_{p3}
\end{aligned}$$

With a simple calculation we can observe that the measure of additive interaction introduced in equation (5.4) is estimated by the parameter β_{p3}. If $\beta_{p3} > 0$ we are in the presence of super-additive interaction between G and E. If $\beta_{p3} < 0$ the interaction is sub-additive. The statistical test associated with the parameter β_{p3} can hence be viewed as a test for additive interaction in the metric of time. Model (5.5) can be extended to include additional covariates, and the interpretation of the product term coefficient as a measure of additive interaction remains valid after conditioning on the additional covariates.

6. Discussion

This doctoral thesis aimed to introduce in the epidemiological literature the percentile approach to the analysis of time-to-event outcomes. The method was first applied on a large prospective cohort of Swedish men and women to assess the association between lifestyle factors and mortality in terms of survival time. These applied studies, which addressed important topics within the field of nutritional epidemiology, were also crucial to explore the advantages of the percentile approach in terms of estimating and presenting results, and provided the root from which the presented methodological developments took place.

This work helped exploiting the advantages of evaluating survival percentiles and provided considerable added value to the fields of nutritional epidemiology and epidemiological methods.

6.1 Advantages of the approach

Studies I-II-III were well received by the scientific community and by the media. Together with the novelty of the findings, three main methodological aspects have largely contributed to improve the quality of these studies: the benefits that presenting time-based measures of association provides to medical research; specific advantages of focusing on survival percentiles; beneficial properties of the adopted statistical method for the estimation of survival percentiles.

Time-based measures of association

Presenting statistical associations with measures that reflect the absolute difference in survival time provides epidemiological studies with considerable advantages in interpreting the results, and facilitates their translation to the general public.

The first advantage of time-based measures of association is that they provide an estimate of the magnitude of the effect, thus overcoming a major limitation of relative measures, such of the HR, described in Section 2.2.4. A measure like the PD conveys information on the excess in survival (i.e. years, days, months) in one group (exposed), as compared to survival in the unexposed group.

Moreover, a measure of exposure-disease association that depicts the difference in survival might be more intuitive and easier to interpret for patients and laymen. Such a measure would provide additional value in clinical and public health decision-making, when individuals are recommended to adhere to a given health recommendation. The notion of *clinical meaningfulness* has been often underlined as essential to understand

treatment effects and statistical associations (Kraemer et al., 2003; Snapinn and Jiang, 2011). Kirk (2001) articulated that *'behavioral scientists are interested in answering three basic questions: first, is an observed result real or should it be attributed to chance? Second, if the result is real, how large is it? Third, is the result large enough to be meaningful and useful?'* Presenting time-based measures of associations, when evaluating time-to-event outcomes, substantially aids results interpretation and facilitates the translation of results from epidemiological studies into relevant public health messages, easing the development of recommendations and public care policies.

The advantages of focusing on a measure that provide a time dimension to the association turned out to be extremely relevant when investigating the context of interaction analysis in Study V. In survival analysis, in fact, no measures of additive interaction, also including information on the magnitude of the interaction effect, were previously available. By focusing on survival percentiles, it was possible to define a measure of additive interaction in the metric of time that represents the excess/decrease in survival due to the presence of both the exposures of interest, thus providing a remarkable contribution to the field.

Survival percentiles

Among the possible methods to calculate time-based measures of association, the survival percentile approach that was herein presented provides specific additional advantages.

First, survival percentiles are perfectly suited to evaluate those situations in which the association of interest is changing over time. By focusing on different percentiles of the observed distribution of events, the association can be evaluated over time, without necessarily assuming a constant effect. A graphical presentation of the associations at various percentiles can be used to assess how this is changing as the proportion of cases increases over time.

Another benefit accrued from estimating survival percentiles is their intuitive interpretation at the individual level. For example, if the 10th survival percentile in a study population is equal to 10 years, we can conclude that 10% of participants experience the event of interest during the first 10 years. By modeling survival percentiles as a function of covariates, these can be predicted for different covariate patterns, and will indicate the time by which participants with specific characteristics achieve a given proportion of events.

Moreover, evaluating survival percentiles allows focusing on the actually observed fraction of cases during follow-up. Presenting differences in median survival may be

appealing because of the simple interpretation of this measure, but it is not common that half of the population has experienced the event of interest during follow-up, and the estimation of median survival would require some data extrapolation. This can be avoided by simply focusing on lower percentiles. This feature is of particular interest in prospective cohort studies when a low fraction of cases is observed, as it typically occurs when dealing with rare outcomes, or when a large fraction of the population is still event-free at the closing of follow-up.

Nevertheless, interpreting specific percentiles such as the 15th estimated in Studies II-III, may not be as straightforward as interpreting the median. Study IV provided a substantial contribution to address this problem. Because of the introduction of delayed entries, a higher number of percentiles is commonly observed when moving the focus from the distribution of survival time to the distribution of attained age. It is pretty common, in prospective cohort studies, that the median age at death is observed even when a low fraction of deaths is documented. Interpreting differences in the median age at event is considerably simpler than interpreting differences in the 15th survival percentile. In Section 5.4.3 results from Study III, originally calculated by focusing on the 15th survival percentiles, were replicated by evaluating median age at death. Results were consistent with the previous, showing the negative effects of red meat at higher values of consumption, but the interpretation of results in terms of differences in median age at death was considerably easier.

Laplace regression

The properties discussed in Section 2.5.3 make Laplace regression a primary choice to model conditional survival percentiles. Laplace offers various advantages in terms of computational speed, estimation performances, and model flexibility. It does not make assumptions required by other methods (such as the one of global linearity) and has shown good performances under different settings and distributional scenarios.

Study IV of this thesis was the first to introduce Laplace regression in the epidemiological literature and to discuss its properties and advantages. This study, which presented the consequences of changing time scale on the interpretation and estimation of the survival function, has improved the potentialities of Laplace regression, extending its possible use to those common situations in which age is of higher interest than follow-up time.

6.2 Added value to nutritional epidemiology

Studies I, II, and III, addressed relevant and controversial topics within the field of nutritional epidemiology.

Study I was the first to evaluate the dose-response relationship between FV consumption and mortality using flexible tools such as splines, and to provide a time-dimension to the association. The methodological strength of the study helped to clarify the effects of low levels of FV intake, which were inconsistent from previous studies. Large part of these inconsistencies might be due to the use of the categorical approach to model the main exposure of FV intake, while it would be reasonable to expect large beneficial effects already at low levels of FV consumption. Moreover, a small number of studies had previously investigated the association between FV consumption and overall mortality, which represents a primary information to understand the health benefits of the nutrients and to plan dietary recommendations. Strengths of the study include the population-based and prospective design, the large sample size, the completeness of ascertainment of deaths through the National Register, and the detailed information on diet. A possible limitation of the study lies in the self-reported nature of the exposure, which can lead to a certain amount of misclassification error.

Study II was the first study to examine the association between sleep duration and survival across levels of total PA, evaluating a suggested hypothesis to explain the negative effects on health of long sleep duration. The lack of association between long sleep and survival among physically active participants seems to support the hypothesis that the previously reported result might be partly explained by comorbidity with low physical activity. This study shared the same population-related strengths of Study I. The main limitation, together with the self-reported nature of the exposure, was the lack of specific questions about sleep quality and factors that could affect the results such as sleep apnea, depression, and employment status.

Study III faced the controversial issue of the health effects of a regular consumption of red meat. Despite the inconsistency of scientific findings, various researchers have started recommending not to eat red meat (Cross et al., 2007; Pan et al., 2012). Results from this study, however, suggest that the association between red meat consumption and mortality might be largely due to the consumption of processed meat. The evaluation of the combined consumption of processed and non-processed meat, in fact, showed that non-processed meat alone was not associated with shorter survival. Together with the strengths described for Study I and Study II, an additional strength of this study was the detailed information on red meat consumption, which allowed dividing the exposure into processed and unprocessed meat.

6.3 Added value to epidemiological methods

Among the public health disciplines, epidemiology is the one where methodology is more relevant and occupies a role of primary importance (Saracci, 1999; Jewell, 2003), and the constant need of methods development in epidemiology has been largely underlined (Rothman et al., 2008). The last 50 years have seen a rapid worldwide growth of epidemiological practice. However, the increased number of epidemiological studies and the improved quality of collected data, have not been always accompanied by the development of novel methods (Pearce and Merletti, 2006).

At the same time, epidemiologists are generally reluctant to accept novel methods, and often prefer to rely on classical approaches, which are as established as they are misused. Weed (2001) has described this common situation as a *mismatch between practice and theory*. This wide gap between theory and practice indicates that newly developed methods can not stand alone, but need to be accompanied by practical tools to facilitate their assimilation, and by numerous practical examples illustrating their use.

This thesis has attempted to move in this direction, easing the crossing between the development of the method and its application. The percentile approach was applied in epidemiological studies to address primary research questions, and the interpretation and advantages of the approach were presented through the use of practical examples. The simultaneous proceeding of the development of the approach and its application in prospective cohort studies may be probably regarded as one of the main strength of this work.

7. Final remarks

7.1 Future research

Nutritional Epidemiology

Results from Studies I-II, and III, have contributed to understand the overall effects on health of lifestyle and dietary factors, but they are far from being the final word on the examined topics. Study I largely confirmed the beneficial contribution of a regular consumption of FV. Future studies might investigate the association between FV and survival from cause-specific mortality, to assess the contribution of concurrent causes of death to the reported results.

Study II showed an interrelationship between sleep duration and PA, suggesting that the negative effects observed at high levels of sleep duration may be attributed to comorbidity with low PA. This, however, remains an hypothesis and requires additional studies to investigate these two lifestyle factors and their interaction in predicting mortality.

Study III concluded that the negative effects of a regular consumption of red meat, reported by different studies, might be largely due to the consumption of processed meat. The conclusion that fresh red meat is not associated with harmful effects, however, is not supported by different researchers, physicians, public health officers, media, and laymen. Further studies on the topic are definitely warranted to fully depict the health-effects of red meat consumption.

Finally, a relevant topic that should be investigated are the health-related effects of combined lifestyle factors. The survival percentile approach could be used to assess the overall difference in survival between participants adhering to different health recommendations and participants who do not meet the recommended thresholds, thus providing an overall estimate, in terms of survival time, of the benefits of an healthy lifestyle.

Epidemiological methods

The studies presented within this doctoral thesis have only made a first step in introducing the percentile approach in the epidemiological literature and in exploiting some of its advantages. A first necessary goal for future research is to increase the regular application of the method, to expand its dissemination and to understand its limitations and the methodological aspects that need further development. Some of the aspects that require further investigation can already be identified.

First, as common in methods dealing with quantiles, an important argument is

to find the optimal tool to present the results. In epidemiology this issue is usually more challenging, as it is common that the evaluation of a single exposure-outcome association requires extensive analyses under different scenarios (e.g. multiple adjustments, different modeling techniques, sensitivity analyses), and various results need to be presented. To fully depict the association between the exposure of interest and the survival distribution, the percentile approach would recommend each model to be evaluated and presented at all observed percentiles, but this option may often turn out to be unfeasible. In Studies I-II-III, for example, it was chosen to limit the results presentation to one single percentile, replicating the only main model at other percentiles as an additional analysis.

Another objective for future research would be to provide a detailed comparison between the percentile approach and other methods that can be used to derive time-based measures of associations in the context of observational epidemiology, discussing advantages and limitations of each method.

The statistical modeling of conditional survival percentiles also deserves additional consideration. Laplace regression has been used throughout this thesis because of the discussed advantages of this method. Nevertheless, the mathematical properties of Laplace should be further investigated, and additional studies are needed to explore its advantages, limitations, and the potentialities of the method in epidemiology. One important issue is that Laplace, at the moment, is only available in the statistical software Stata. The development of a SAS procedure and a R package to implement the method would provide a considerably larger number of researchers to use this tool. Finally, it should be stressed that all the presented advantages of evaluating survival percentiles in epidemiology are not conditioned on using Laplace regression. Further development of other methods to model conditional percentiles of censored outcomes would certainly improve the potentialities of the percentile approach.

One important issue that has received little attention is the application and meaning of the percentile approach when investigating specific outcomes whose occurrence is not certain. Extensions of the method to other endpoints rather than all-cause mortality does not represent a substantial methodological threat, but makes interpretation of results less intuitive. In addition, when investigating the time to developing specific conditions, one may need to take into account the presence of competing events. A framework to evaluate survival percentiles in a competing-risk setting has never been investigated.

When focusing on survival percentile, statistical associations could be also presented in terms of percentile ratios (Uno et al., 2014), which provide an estimate of

the exposure-outcome association at the pth survival percentiles in relative terms. Another interesting aspect that future research might address is to use the properties of the quantiles to develop a regression-based framework for estimating multivariable-adjusted percentile ratios.

7.2 Conclusion

Our findings have provided significant contributions to understand the association between lifestyle factors and survival:

- A daily consumption of 5 servings/day of fruit and vegetables is associated with the longest survival. Lower levels of consumption are associated with shorter survival up to 3 years for individuals who do not consume fruit and vegetables (Study I).
- Sleep duration below and above 7 hours per day are associated with progressively shorter survival. However, the association between long sleep duration and survival might be partly explain by comorbidity between long sleep duration and a low level of physical activity (Study II).
- Higher levels of red meat consumption are associated with shorter survival. However, the negative effects are largely due to the consumption of processed red meat, as unprocessed red meat alone was not associated with shorter survival (Study III).

The methodological studies have contributed to the development of the percentile approach to time-to-event outcomes and have presented some of its advantages:

- The percentile approach provides a flexible method to evaluate the survival distribution, and to understand the link between the risk of experiencing an event of interest and the time by which this is attained. The introduction of a statistical technique to estimate conditional survival percentiles, and its extension to evaluate percentiles of attained age at the event, has substantially enriched its potentialities and allowed its application in epidemiological research (Study IV).
- Evaluating survival percentiles substantially contributes to the analysis of interaction, as it allows deriving a measure of additive interaction in the metric of time (Study V).

Appendix

Appendix I - Details on Laplace regression

Laplace regression aims to estimate the τth conditional quantile of T_i. A more general form of the model introduced in (2.12) can be defined by also including the scale parameter $\sigma(\tau)$:

$$t_i = \mathbf{x}_i^T \boldsymbol{\beta}(\tau) + \sigma_i(\tau) u_i \tag{7.1}$$

Conditionally on $\mathbf{x}_i$, T_i is assumed to follow an AL distribution with probability density function

$$f(t_i|\mathbf{x}_i, \sigma, \tau) = \exp\left[I(t_i \leq \mathbf{x}_i^T \boldsymbol{\beta}(\tau) - \tau)\frac{t_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)}{\sigma(\tau)}\right]\frac{\tau(1-\tau)}{\sigma(\tau)} \tag{7.2}$$

and cumulative distribution function

$$F(t_i|\mathbf{x}_i, \sigma, \tau) = \exp\left[I(t_i \leq \mathbf{x}_i^T \boldsymbol{\beta}(\tau) - \tau)\frac{t_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)}{\sigma(\tau)}\right](\tau - I(t_i > \mathbf{x}_i^T \boldsymbol{\beta}(\tau))) + I(t_i > \mathbf{x}_i^T \boldsymbol{\beta}(\tau)) \tag{7.3}$$

In the presence of censored observations, Y_i is observed in the place of T_i and the log-likelihood function, based on (7.2) and (7.3), can be written as proportional to

$$\begin{aligned} l_n\{\boldsymbol{\beta}(\tau), \sigma(\tau)|y_i, \mathbf{x}_i, \delta_i\} = \sum_{i=1}^{n} \delta_i \left\{(\omega_i - p)\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)}{\sigma(\tau)} + \log\frac{\tau(1-\tau)}{\sigma(\tau)}\right\} \\ + (1-\delta_i)\omega_i \log\left[1 - \tau\exp\left\{(1-\tau)\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)}{\sigma(\tau)}\right\}\right] \\ + (1-\delta_i)(1-\omega_i)\left\{\log(1-\tau) - \tau\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)}{\sigma(\tau)}\right\} \end{aligned} \tag{7.4}$$

where $\delta_i = I(t_i \leq c_i)$ and $\omega_i = I(y_i \leq \mathbf{x}_i^T \boldsymbol{\beta}(\tau))$

Estimation is then performed by maximizing the first derivatives of $l_n\{\boldsymbol{\beta}(\tau), \sigma(\tau)|y_i, \mathbf{x}_i, \delta_i\}$ with respect to $\boldsymbol{\beta}(\tau)$ and $\sigma(\tau)$, which are

$$S_n\{\boldsymbol{\beta}(\tau)\} = \frac{1}{\sigma(\tau)}\sum_{i=1}^{n} \mathbf{x}_i \left\{\tau - \omega_i - \omega_i(1-\delta_i)\frac{\tau - 1}{1 - F(y_i|\mathbf{x}_i)}\right\} \tag{7.5}$$

$$S_n\{\sigma(\tau)\} = \frac{1}{\sigma(\tau)} \sum_{i=1}^{n} \left[\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}(\tau)}{\sigma(\tau)} \left\{ \tau - \omega_i - \omega_i(1-\delta_i) \frac{\tau - 1}{1 - F(y_i|\mathbf{x}_i)} \right\} - \delta_i \right] \quad (7.6)$$

These equations do not have a closed form solution and an iterative procedure can be adopted. The original paper introducing Laplace regression suggested to use an algorithm proposed by Nelder and Mead (Nelder and Mead, 1965), and to complete inference on the parameters by bootstrap. Recently, a gradient search algorithm for estimating quantiles has been introduced and is currently adopted for the estimation of Laplace regression (Bottai et al., 2015).

This algorithm, similarly to the Newton-Raphson, is based on the gradient of the log-likelihood that generates a finite sequence of parameters values along which the likelihood increases. It has a wide applicability, and it may be regarded as a general algorithm for unconstrained optimization of any objective continuous, concave, and first-order differentiable function. The algorithm has shown a remarkable computational speed, through a large simulation study.

Specifically, based on the Laplace-likelihood defined in 7.4, we can defined the gradient

$$g(\boldsymbol{\beta}(\tau)) = -\sum_{i=1}^{n} \mathbf{x}_i (I_{y_i \leq \mathbf{x}_i^T \boldsymbol{\beta}(\tau)} - \tau) \quad (7.7)$$

that has the property of being equal to the first derivative of l_n with respect to $\boldsymbol{\beta}(\tau)$ at all points of the parameter space. The algorithm searches the positive semi-line in the direction of $g(\boldsymbol{\beta}(\tau))$ for a new parameter value at which the likelihood is larger, stopping when the change in the log-likelihood is below a pre-specified tolerance threshold.

Koenker (2011) raised some concerns about the performances of Laplace regression, which were largely addressed by the authors of the original paper (Bottai and Zhang, 2011). Laplace regression, like the majority of methods dealing with censored data, presents a small bias. However, the method has a low mean squared error and different studies have consistently showed its good performances in estimating regression parameters under different simulation scenarios (Bottai and Zhang, 2010; Bottai and Orsini, 2013; Bottai et al., 2015; Bellavia et al., 2015b). Laplace is, at the moment, the best available technique to estimate conditional survival percentiles. Future methodological studies are nonetheless recommended to optimize its performances and deeply explore its mathematical properties.

Appendix II - Stata Tutorial

The command `laplace` to fit Laplace regression models in Stata can be downloaded and installed with the following code:

```
net install laplace, from(http://www.imm.ki.se/biostatistics/stata) replace
```

To illustrate the use of the command the dataset `kidney_ca` will be used. This dataset includes data from a clinical trial on 347 patients with metastatic renal carcinoma, who were assigned to standard (interferon-α - IFN) or novel (oral medroxyprogesterone - MPA) treatment. The dataset can be downloaded and opened in Stata by typing:

```
use http://www.imm.ki.se/biostatistics/data/kidney_ca, clear
```

Dichotomous exposures

The main variables of the dataset are `survtime`, which represents the time to event or censoring, in days, the failure status `cens` (0 = censored, 1 = death), and `trt`, a binary variable indicating the assigned treatment. A total of 322 patients died during follow-up (93%), so that it is possible to estimate, without data extrapolation, all survival percentiles from the 1st to the 93th.

First, we fit Laplace regression models on the 25th, 50th, and 75th percentile of survival, without including any covariate in the model.

$$T(p) = \beta_{p0} \tag{7.8}$$

with p=(0.25, 0.50, 0.75).

These three models can be fitted in Stata by one single fit. Laplace models are estimated by writing the command `laplace` followed by the outcome and eventual covariates (none in this first example). The failure information and the desired quantiles must be indicated among the options. The same results of this crude models can be obtained by calculating survival percentiles from the Kaplan-Meier estimates of the survival function (command `stci`, also reported).

[1]. qui: stset survtime, failure(cens) scale(30.4)

. laplace _t, failure(cens) quantile(.25 .5 .75)

Laplace regression — No. of subjects = 347; No. of failures = 322

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
q25						
_cons	3.05922	.3233314	9.46	0.000	2.425502	3.692938
q50						
_cons	7.960537	.6174112	12.89	0.000	6.750433	9.17064
q75						
_cons	16.51316	1.156959	14.27	0.000	14.24556	18.78075

. stci, p(25)

failure _d: cens
analysis time _t: survtime/30.4

	no. of subjects	25%	Std. Err.	[95% Conf.	Interval]
total	347	3.059211	.3939976	2.30263	4.01316

. stci, p(50)

failure _d: cens
analysis time _t: survtime/30.4

	no. of subjects	50%	Std. Err.	[95% Conf.	Interval]
total	347	7.960526	.5702351	6.90789	9.17763

. stci, p(75)

failure _d: cens
analysis time _t: survtime/30.4

	no. of subjects	75%	Std. Err.	[95% Conf.	Interval]
total	347	16.51316	1.410827	15.0658	20.8882

The first 25% of the study population experienced the event of interest after 3 months. Median survival was equal to 8 months. After 16.5 months 75% of the population has experienced the event of interest while 25% was still event-free.

[1]The `stset` command is included to change the unit of the time variable from days to months.

We now model survival percentiles as a function of the assigned treatment, to estimate PDs across treatment groups at the 25th, 50th, and 75th percentile.

$$T(p|trt) = \beta_{p0} + \beta_{p1} \cdot trt \tag{7.9}$$

with; p=(0.25, 0.50, 0.75). In Stata:

```
. laplace _t trt, failure(cens) quantile(.25 .5 .75)

Laplace regression                              No. of subjects  =        347
                                                No. of failures  =        322
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
q25						
trt	1.477664	.7145055	2.07	0.039	.0772591	2.878069
_cons	2.500004	.3734526	6.69	0.000	1.768051	3.231958
q50						
trt	3.107737	1.220632	2.55	0.011	.7153415	5.500132
_cons	6.809208	.7192481	9.47	0.000	5.399508	8.218909
q75						
trt	3.813401	4.01896	0.95	0.343	-4.063616	11.69042
_cons	15.88817	1.605474	9.90	0.000	12.7415	19.03484

The first 25% of events occur 1.5 months later in the treated group, while median survival was delayed by 3 months.

This analysis can be extended by further adjusting for additional covariates. For example, the dataset include information on baseline age (variable `age`), which can be included in the model to estimate age-adjusted differences in survival percentiles:

$$T(p|trt, age) = \beta_{p0} + \beta_{p1} \cdot trt + \beta_{p2} \cdot age \tag{7.10}$$

p=(0.25, 0.50, 0.75). In Stata:

```
. laplace _t trt age, failure(cens) quantile(.25 .5 .75)

Laplace regression                              No. of subjects  =        347
                                                No. of failures  =        322
```

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
q25						
trt	1.60993	.7489187	2.15	0.032	.142076	3.077783
age	.0222974	.0350038	0.64	0.524	-.0463087	.0909035
_cons	1.164572	2.088221	0.56	0.577	-2.928266	5.25741
q50						
trt	3.046283	1.218155	2.50	0.012	.6587419	5.433824
age	-.0090226	.0627681	-0.14	0.886	-.1320458	.1140005
_cons	7.339	3.686768	1.99	0.047	.113067	14.56493
q75						
trt	3.816384	4.261885	0.90	0.371	-4.536758	12.16953
age	-.0185994	.1421905	-0.13	0.896	-.2972876	.2600888
_cons	16.95907	8.946016	1.90	0.058	-.5747974	34.49294

Adjusting for baseline age did not substantially change the treatment effect.

Continuous exposures

Inclusion of a continuous exposure in the model will implicitly assume that the relationships between the predictor and the estimated survival percentiles are all linear. To relax this assumption and test for possible non-linear associations we can serve of spline transformations with the Stata command `mkspline` and plot the dose-response with the command `xblc` (Orsini and Greenland, 2011).

The kidney dataset includes information on the continuous predictor of blood hemoglobin content (g/dl, variable `haem`), which is a possible predictor of survival. The next Stata window shows how to estimate the age-adjusted association between hemoglobin level and median survival (50th percentile), flexibly modeling the exposure with restricted cubic splines, and plotting results with the median value of the covariate as reference.

```
. mkspline haems = haem , nk(3) cubic display
```

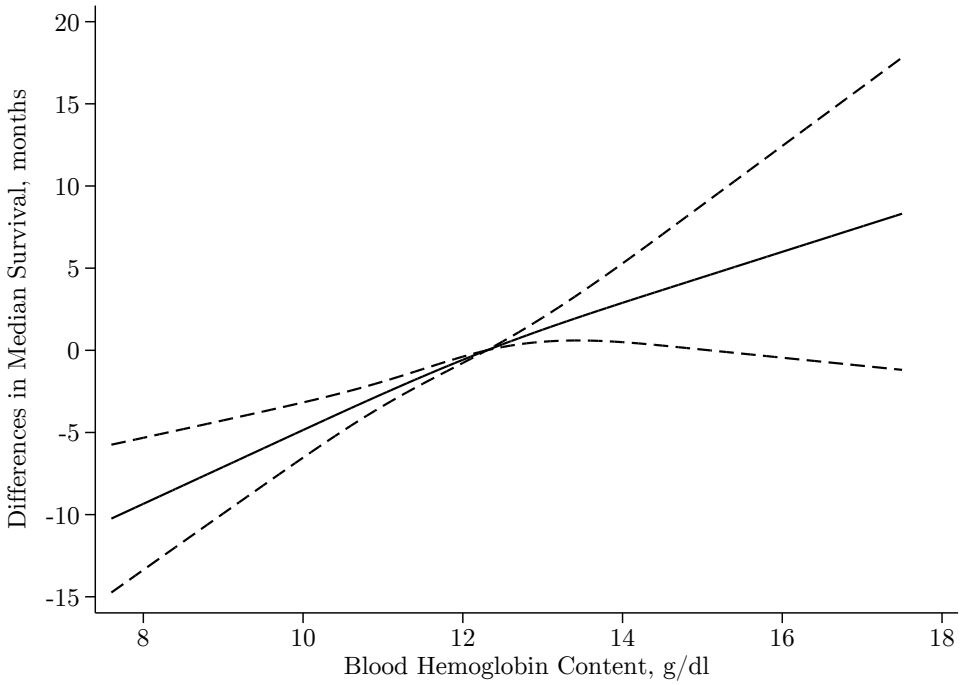
	knot1	knot2	knot3
haem	9.7	12.3	14.7

```
. laplace _t haems* age, failure(cens) quantile(.5)
```

Laplace regression — No. of subjects = 347; No. of failures = 322

_t	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
q50						
haems1	2.246168	.6158796	3.65	0.000	1.039067	3.45327
haems2	-.4479796	.994397	-0.45	0.652	-2.396962	1.501003
age	.0534481	.0622114	0.86	0.390	-.068484	.1753801
_cons	-21.01797	8.240127	-2.55	0.011	-37.16833	-4.867622

```
. qui levelsof haem
. qui xblc haems*, cov(haem) ref(12.3) at(`r(levels)´) line
```



The graph shows that higher hemoglobin content was approximately linearly associated with progressively longer median survival.

Acknowledgments

The work of these years was possible thanks to the support and contribution of different people.

First, I would like to thank **Nicola Orsini**, my main supervisor, for giving me the possibility of doing this PhD. Thanks for the continuous support and constant interaction, and thanks for having made possible such a friendly working environment. These years have been fantastic. And of course, thanks for the *quotes* you shared with us, which we will jealously treasure.

A special thanks to my co-supervisors, **Alicja Wolk** and **Matteo Bottai**. Alicja, thanks for introducing me to the challenging world of nutritional epidemiology, and for sharing your extensive experience with helpful comments. Thanks also for having me involved in different collaborations within our unit: I have learned a lot. (Also, thanks for trying to teach me Swedish, but *jag är en katastrof*). Matteo, thank you for the time you gave me in insightful discussions. This work would have not been possible without your constant supervision and suggestions. And, of course, without Laplace regression!

Taking a step back in time, I wish to express my sincere gratitude to **Giorgio Vittadini, Giovanni Corrao** and **Andrea Baccarelli**. If I started this beautiful adventure it is only because I had the fortune to meet these exceptional teachers, who through their passion for research introduced me to this fascinating world.

Thanks also to the coauthors of the studies included in this thesis, **Susanne Larsson, Torbjörn Åkerstedt, Andrea Discacciati**, for their important contribution, and to **Paolo Frumento** for the essential and constant help with Laplace regression.

A particular thanks to **Rino Bellocco** for involving me in the organization of the Summer School and in other collaborations, and for the many engaging and broad discussions.

Thanks to all my colleagues at the Unit of Nutritional Epidemiology. **Andrea Discacciati** and **Alessio Crippa**, **Viktor Oskarsson**, **Daniela Di Giuseppe**. **Alice Wallin**, **Susanne Rautiainen Lagerström**, **Agneta Åkesson, Jinjin Zheng Selin, Mimi Throne-Holst, Otto Stackelberg, Niclas Håkansson, Thanasis Tektonidis, Frej Stilling**, and **Nada Hana**. Thanks also to **Joana Kaluza** and **Carolina Donat Vargas**, and to all my past colleagues: **Iffat Rahman, Holly Harris, Camilla Olofsson, Laura Thomas, Becky Leung, Ann Burgaz, Katica Anusic, Prapasri Ljungberg, Charlotte Bergqvist, Lollo Sjöholm, Anna Ingemarsdotter, Bettina Julin**.

Thanks also to the colleagues at the Unit of Biostatistics (**Silvia Columbu, Davide Bossoli, Celia Garcia Pareja, Xin Fang, Daniel Olsson, Michele Santacatterina**) and all the many other colleagues at IMM and KI I have met over the years.

Last but not least, thanks to all my family and friends in Italy, Sweden, UK, and US. A particular thanks to **Paolo** for the useful comments on the draft of this thesis.

Bibliography

Agudo, A., Cabrera, L., Amiano, P., Ardanaz, E., Barricarte, A., Berenguer, T., Chirlaque, M. D., Dorronsoro, M., Jakszyn, P., Larrañaga, N., Martínez, C., Navarro, C., Quirós, J. R., Sánchez, M. J., Tormo, M. J., and González, C. A. (2007). Fruit and vegetable intakes, dietary antioxidant nutrients, and total mortality in Spanish adults: findings from the Spanish cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-Spain). *The American Journal of Clinical Nutrition*, 85(6):1634–1642.

Altman, D. G. (1990). *Practical Statistics for Medical Research*. CRC Press.

Andersen, P. K., Geskus, R. B., Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41(3):861–870.

Austin, P. C. and Schuster, T. (2014). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. *Statistical Methods in Medical Research*.

Bang, H. and Tsiatis, A. A. (2002). Median regression with censored cost data. *Biometrics*, 58(3):643–649.

Bazzano, L. A., Serdula, M. K., and Liu, S. (2003). Dietary intake of fruits and vegetables and risk of cardiovascular disease. *Current atherosclerosis reports*, 5(6):492–499.

Bellavia, A., Bottai, M., Discacciati, A., and Orsini, N. (2015a). Adjusted survival curves with multivariable laplace regression. *Epidemiology*, 26(2):e17–18.

Bellavia, A., Bottai, M., and Orsini, N. (2016). Evaluating additive interaction using survival percentiles. *Epidemiology*, In press.

Bellavia, A., Bottai, M., Wolk, A., and Orsini, N. (2013a). Physical activity and mortality in a prospective cohort of middle-aged and elderly men - a time perspective. *The International Journal of Behavioral Nutrition and Physical Activity*, 10:94.

Bellavia, A., Bottai, M., Wolk, A., and Orsini, N. (2014a). Alcohol consumption and mortality: a dose-response analysis in terms of time. *Annals of Epidemiology*, 24(4):291–296.

Bellavia, A., Discacciati, A., Bottai, M., Wolk, A., and Orsini, N. (2015b). Using Laplace Regression to Model and Predict Percentiles of Age at Death When Age Is the Primary Time Scale. *American Journal of Epidemiology*, 182(3):271–277.

Bellavia, A., Åkerstedt, T., Bottai, M., Wolk, A., and Orsini, N. (2014b). Sleep duration and survival percentiles across categories of physical activity. *American Journal of Epidemiology*, 179(4):484–491.

Bellavia, A., Larsson, S. C., Bottai, M., Wolk, A., and Orsini, N. (2013b). Fruit and vegetable consumption and all-cause mortality: a dose-response analysis. *The American Journal of Clinical Nutrition*, 98(2):454–459.

Bellavia, A., Larsson, S. C., Bottai, M., Wolk, A., and Orsini, N. (2014c). Differences in survival associated with processed and with nonprocessed red meat consumption. *The American Journal of Clinical Nutrition*, 100(3):924–929.

Bellavia, A., Wolk, A., and Orsini, N. (2015c). Differences in age at death according to smoking and age at menopause. *Menopause*.

Beyerlein, A. (2014). Quantile regression—opportunities and challenges from a user's perspective. *American journal of epidemiology*, 180(3):330–331.

Bobbio, M., Demichelis, B., and Giustetto, G. (1994). Completeness of reporting trial results: effect on physicians' willingness to prescribe. *The Lancet*, 343(8907):1209–1211.

Boffetta, P., Couto, E., Wichmann, J., Ferrari, P., Trichopoulos, D., Bueno-de Mesquita, H. B., Van Duijnhoven, F. J., Büchner, F. L., Key, T., Boeing, H., and others (2010). Fruit and vegetable intake and overall cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Journal of the National Cancer Institute*, 102(8):529–537.

Boscovich, R. J. (1757). De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. *Bononiensi Scientarum et Artum Instituto Atque Academia Commentarii*, 4:353–396.

Bottai, M., Cai, B., and McKeown, R. E. (2010). Logistic quantile regression for bounded outcomes. *Statistics in medicine*, 29(2):309–317.

Bottai, M. and Orsini, N. (2013). A command for Laplace regression. *Stata J*, 13(2):1–13.

Bottai, M., Orsini, N., and Geraci, M. (2015). A gradient search maximization algorithm for the asymmetric Laplace likelihood. *Journal of Statistical Computation and Simulation*, 85(10):1919–1925.

Bottai, M. and Zhang, J. (2010). Laplace regression with censored data. *Biometrical Journal*, 52(4):487–503.

Bottai, M. and Zhang, J. (2011). Authors' reply. *Biometrical Journal*, 53(5):861–866.

Botto, L. D. and Khoury, M. J. (2001). Commentary: Facing the Challenge of Gene-Environment Interaction: The Two-by-Four Table and Beyond. *American Journal of Epidemiology*, 153(10):1016–1020.

Brookmeyer, R. and Crowley, J. (1982). A Confidence Interval for the Median Survival Time. *Biometrics*, 38(1):29–41.

Cappuccio, F. P., D'Elia, L., Strazzullo, P., and Miller, M. A. (2010). Sleep Duration and All-Cause Mortality: A Systematic Review and Meta-Analysis of Prospective Studies. *Sleep*, 33(5):585–592.

Chan, D. S. M., Lau, R., Aune, D., Vieira, R., Greenwood, D. C., Kampman, E., and Norat, T. (2011). Red and Processed Meat and Colorectal Cancer Incidence: Meta-Analysis of Prospective Studies. *PLoS ONE*, 6(6):e20456.

Cheung, Y. B., Gao, F., and Khoo, K. S. (2003). Age at diagnosis and the choice of survival analysis methods in cancer epidemiology. *Journal of Clinical Epidemiology*, 56(1):38–43.

Cologne, J., Hsu, W.-L., Abbott, R. D., Ohishi, W., Grant, E. J., Fujiwara, S., and Cullings, H. M. (2012). Proportional Hazards Regression in Epidemiologic Follow-up Studies: An Intuitive Consideration of Primary Time Scale. *Epidemiology*, 23(4):565–573.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cross, A. J., Leitzmann, M. F., Gail, M. H., Hollenbeck, A. R., Schatzkin, A., and Sinha, R. (2007). A Prospective Study of Red and Processed Meat Intake in Relation to Cancer Risk. *PLoS Med*, 4(12):e325.

Dauchet, L., Amouyel, P., Hercberg, S., and Dallongeville, J. (2006). Fruit and vegetable consumption and risk of coronary heart disease: a meta-analysis of cohort studies. *The Journal of nutrition*, 136(10):2588–2593.

Durrleman, S. and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press.

Farcomeni, A. (2010). Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Statistics and Computing*, 22(1):141–152.

Forrow, L., Taylor, W. C., and Arnold, R. M. (1992). Absolutely relative: How research results are summarized can affect treatment decisions. *The American Journal of Medicine*, 92(2):121–124.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (2010). Survival Analysis. In *Fundamentals of Clinical Trials*, pages 269–291. Springer New York.

Frumento, P. and Bottai, M. (2015). Parametric modeling of quantile regression coefficient functions. *Biometrics*.

Gangwisch, J. E., Heymsfield, S. B., Boden-Albala, B., Buijs, R. M., Kreier, F., Opler, M. G., Pickering, T. G., Rundle, A. G., Zammit, G. K., and Malaspina, D. (2008). Sleep Duration Associated with Mortality in Elderly, but not Middle-Aged, Adults in a Large US Sample. *Sleep*, 31(8):1087–1096.

Genkinger, J. M., Platz, E. A., Hoffman, S. C., Comstock, G. W., and Helzlsouer, K. J. (2004). Fruit, Vegetable, and Antioxidant Intake and All-Cause, Cancer, and Cardiovascular Disease Mortality in a Community-dwelling Population in Washington County, Maryland. *American Journal of Epidemiology*, 160(12):1223–1233.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8(1):140–154.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24(3):461–479.

Ghali, W., Quan, H., Brant, R., Van Melle, G., Norris, C., Faris, P., Galbraith, D., and Knudtson, M. (2001). Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models. *JAMA*, 286(12):1494–1497.

Global Burden of Disease (2015). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*.

Greenland, S. (1983). Tests for interaction in epidemiologic studies: A review and a study of power. *Statistics in Medicine*, 2(2):243–251.

Greenland, S. (1993). Basic problems in interaction assessment. *Environmental Health Perspectives*, 101(Suppl 4):59–66.

Greenland, S. (1995a). Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology*, 6(4):450–454.

Greenland, S. (1995b). Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. *Epidemiology*, 6(4):356–365.

Greenland, S. (1995c). Problems in the Average-Risk Interpretation of Categorical Dose-Response Analyses. *Epidemiology*, 6(5):563–565.

Greenland, S. (2009). Interactions in Epidemiology: Relevance, Identification, and Estimation:. *Epidemiology*, 20(1):14–17.

Hansson, L. M. and Galanti, M. R. (2000). Diet-associated risks of disease and self-reported food consumption: how shall we treat partial nonresponse in a food frequency questionnaire? *Nutrition and cancer*, 36(1):1–6.

Harris, H., Hakansson, N., Olofsson, C., Stackelberg, O., Julin, B., Åkesson, A., and Wolk, A. (2013). The Swedish mammography cohort and the cohort of Swedish men: study design and characteristics of two population-based longitudinal cohorts. *2013*, (2):1:16.

Hernán, M. A. (2010). The Hazards of Hazard Ratios. *Epidemiology*, 21(1):13–15.

Hosmer, D., Lemeshow, S., and May, S. (2011). *Applied Survival Analysis: Regression Modeling of Time to Event Data, 2nd Edition*.

Hublin, C., Partinen, M., Koskenvuo, M., and Kaprio, J. (2007). Sleep and Mortality: A Population-Based 22-Year Follow-Up Study. *Sleep*, 30(10):1245–1253.

Hung, H.-C., Joshipura, K. J., Jiang, R., Hu, F. B., Hunter, D., Smith-Warner, S. A., Colditz, G. A., Rosner, B., Spiegelman, D., and Willett, W. C. (2004). Fruit and

Vegetable Intake and Risk of Major Chronic Disease. *Journal of the National Cancer Institute*, 96(21):1577–1584.

Hux, J. E. and Naylor, C. D. (1995). Communicating the Benefits of Chronic Preventive Therapy Does the Format of Efficacy Data Determine Patients' Acceptance of Treatment? *Medical Decision Making*, 15(2):152–157.

Jewell, N. (2003). *Statistics for Epidemiology*. Chapman and Hall.

Kalbfleisch, J. D. and Prentice, R. L. (1981). Estimation of the average hazard ratio. *Biometrika*, 68(1):105–112.

Kaluza, J., Wolk, A., and Larsson, S. C. (2012). Red Meat Consumption and Risk of Stroke A Meta-Analysis of Prospective Studies. *Stroke*, 43(10):2556–2560.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kappeler, R., Eichholzer, M., and Rohrmann, S. (2013). Meat consumption and diet quality and mortality in NHANES III. *European Journal of Clinical Nutrition*, 67(6):598–606.

Khoury, M. J., Gwinn, M., and Ioannidis, J. P. A. (2010). The Emergence of Translational Epidemiology: From Scientific Discovery to Population Health Impact. *American Journal of Epidemiology*, 172(5):517–524.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61(2):213–218.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis*. Statistics for Biology and Health. Springer New York, New York, NY.

Knol, M. J., Egger, M., Scott, P., Geerlings, M. I., and Vandenbroucke, J. P. (2009). When One Depends on the Other: Reporting of Interaction in Case-Control and Cohort Studies. *Epidemiology*, 20(2):161–166.

Knol, M. J. and VanderWeele, T. J. (2012). Recommendations for presenting analyses of effect modification and interaction. *International Journal of Epidemiology*, 41(2):514–520.

Knol, M. J., VanderWeele, T. J., Groenwold, R. H. H., Klungel, O. H., Rovers, M. M., and Grobbee, D. E. (2011). Estimating measures of interaction on an additive scale for preventive exposures. *European Journal of Epidemiology*, 26(6):433–438.

Knutson, K. and Turek, T. (2006). The U-shaped association between sleep and health: the 2 peaks do not mean the same thing. *Sleep*, 29(7):878–879.

Koenker, R. (2005). *Quantile regression*. Cambridge university press.

Koenker, R. (2008). Censored quantile regression redux. *Journal of Statistical Software*, 27(6):1–25.

Koenker, R. (2011). "A note on Laplace regression with censored data". *Biometrical Journal. Biometrische Zeitschrift*, 53(5):855–860; author reply 861–866.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468.

Koenker, R. and Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310.

Kom, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-Event Analysis of Longitudinal Follow-up of a Survey: Choice of the Time-scale. *American Journal of Epidemiology*, 145(1):72–80.

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., and Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child & Adolescent Psychiatry*, 42(12):1524–1529.

Kripke, D. F., Garfinkel, L., Wingard, D. L., Klauber, M. R., and Marler, M. R. (2002). Mortality associated with sleep duration and insomnia. *Archives of General Psychiatry*, 59(2):131–136.

Lai, T. and Su, Z. (2006). Confidence intervals for survival quantiles in the cox regression model. *Lifetime Data Analysis*, 12(4):407-419.

Lamarca, R., Alonso, J., Gómez, G., and Muñoz, l. (1998). Left-truncated Data With Age as Time Scale: An Alternative for Survival Analysis in the Elderly Population. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 53A(5):M337–M343.

Larsson, S. C. and Orsini, N. (2013). Red Meat and Processed Meat Consumption and All-Cause Mortality: A Meta-Analysis. *American Journal of Epidemiology*.

Larsson, S. C., Orsini, N., and Wolk, A. (2006). Processed Meat Consumption and Stomach Cancer Risk: A Meta-Analysis. *Journal of the National Cancer Institute*, 98(15):1078–1087.

Lauretani, F., Semba, R. D., Dayhoff-Brannigan, M., Corsi, A. M., Iorio, A. D., Buiatti, E., Bandinelli, S., Guralnik, J. M., and Ferrucci, L. (2008). Low total plasma carotenoids are independent predictors of mortality among older persons. *European Journal of Nutrition*, 47(6):335–340.

Lawless, J. F. (2011). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons.

Lee, D. and Neocleous, T. (2010). Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(5):905–920.

Li, R. and Chambless, L. (2007). Test for Additive Interaction in Proportional Hazards Models. *Annals of Epidemiology*, 17(3):227–236.

Liestol, K. and Andersen, P. K. (2002). Updating of covariates and choice of time origin in survival analysis: problems with vaguely defined disease states. *Statistics in Medicine*, 21(23):3701–3714.

Liu, Y. and Bottai, M. (2009). Mixed-Effects Models for Conditional Quantiles with Longitudinal Data. *The International Journal of Biostatistics*, 5(1).

Ludvigsson, J. F., Otterblad-Olausson, P., Pettersson, B. U., and Ekbom, A. (2009). The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *European Journal of Epidemiology*, 24(11):659–667.

Lytsy, P., Berglund, L., and Sundström, J. (2012). A proposal for an additional clinical trial outcome measure assessing preventive effect as delay of events. *European Journal of Epidemiology*, 27(12):903–909.

Mackenzie, T. (2012). Survival Curve Estimation with Dependent Left Truncated Data Using Cox's Model. *The International Journal of Biostatistics*, 8(1).

Marmot, M., Atinmo, T., Byers, T., Chen, J., Hirohata, T., Jackson, A., James, W., Kolonel, L., Kumanyika, S., Leitzmann, C., Mann, J., Powers, H., Reddy, K.,

Riboli, E., Rivera, J. A., Schatzkin, A., Seidell, J., Shuker, D., Uauy, R., Willett, W., and Zeisel, S. (2007). Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective. Report, World Cancer Research Fund / American Institute for Cancer Research, Washington DC, US.

Marrie, R. A., Dawson, N. V., and Garland, A. (2009). Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. *Journal of Clinical Epidemiology*, 62(5):511–517.e1.

Micha, R., Michas, G., and Mozaffarian, D. (2012). Unprocessed Red and Processed Meats and Risk of Coronary Artery Disease and Type 2 Diabetes – An Updated Review of the Evidence. *Current Atherosclerosis Reports*, 14(6):515–524.

Michels, K. B. (2003). Nutritional epidemiology—past, present, future. *International Journal of Epidemiology*, 32(4):486–488.

Naylor, C. D., Chen, E., and Strauss, B. (1992). Measured Enthusiasm: Does the Method of Reporting Trial Results Alter Perceptions of Therapeutic Effectiveness? *Annals of Internal Medicine*, 117(11):916–921.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313.

Nelson, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*, 14(4):945–966.

Nicklett, E. J., Semba, R. D., Xue, Q.-L., Tian, J., Sun, K., Cappola, A. R., Simonsick, E. M., Ferrucci, L., and Fried, L. P. (2012). Fruit and Vegetable Intake, Physical Activity, and Mortality in Older Community-Dwelling Women. *Journal of the American Geriatrics Society*, 60(5):862–868.

Nieto, F. J. and Coresh, J. (1996). Adjusting Survival Curves for Confounders: A Review and a New Method. *American Journal of Epidemiology*, 143(10):1059–1068.

Norman, A., Bellocco, R., Bergstrom, A., and Wolk, A. (2001). Validity and reproducibility of self-reported total physical activity: differences by relative weight. *International Journal of Obesity and Related Metabolic Disorders*, 25:582-8.

Orsini, N., Bellocco, R., Bottai, M., Hagströmer, M., Sjöström, M., Pagano, M., and Wolk, A. (2008). Validity of self-reported total physical activity questionnaire among older women. *European Journal of Epidemiology*, 23(10):661–667.

Orsini, N. and Greenland, S. (2011). A procedure to tabulate and plot results after flexible modeling of a quantitative covariate. *Stata Journal*, 11(1):1.

Orsini, N., Wolk, A., and Bottai, M. (2012). Evaluating percentiles of survival. *Epidemiology*, 23(5):770–771.

Pan, A., Sun, Q., Bernstein, A., Schulze, M., Manson, J. E., Stampfer, M., Willett, W., and Hu, F. B. (2012). Red meat consumption and mortality: Results from 2 prospective cohort studies. *Archives of Internal Medicine*, 172(7):555–563.

Pearce, N. and Merletti, F. (2006). Complexity, simplicity, and epidemiology. *International Journal of Epidemiology*, 35(3):515–519.

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482).

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012.

Powell, J. L. (1986). Censored regression quantiles. *Journal of econometrics*, 32(1):143–155.

Ray, A. L., Semba, R. D., Walston, J., Ferrucci, L., Cappola, A. R., Ricks, M. O., Xue, Q.-L., and Fried, L. P. (2006). Low Serum Selenium and Total Carotenoids Predict Mortality among Older Women Living in the Community: The Women's Health and Aging Studies. *The Journal of Nutrition*, 136(1):172–176.

Rissanen, T. H., Voutilainen, S., Virtanen, J. K., Venho, B., Vanharanta, M., Mursu, J., and Salonen, J. T. (2003). Low Intake of Fruits, Berries and Vegetables Is Associated with Excess Mortality in Men: the Kuopio Ischaemic Heart Disease Risk Factor (KIHD) Study. *The Journal of Nutrition*, 133(1):199–204.

Rohrmann, S., Overvad, K., Bueno-de Mesquita, H. B., Jakobsen, M. U., Egeberg, R., Tjønneland, A., Nailler, L., Boutron-Ruault, M.-C., Clavel-Chapelon, F., Krogh, V., Palli, D., Panico, S., Tumino, R., Ricceri, F., Bergmann, M. M., Boeing, H., Li, K., Kaaks, R., Khaw, K.-T., Wareham, N. J., Crowe, F. L., Key, T. J., Naska, A., Trichopoulou, A., Trichopoulos, D., Leenders, M., Peeters, P. H., Engeset, D., Parr, C. L., Skeie, G., Jakszyn, P., Sánchez, M.-J., Huerta, J. M., Redondo, M. L., Barricarte, A., Amiano, P., Drake, I., Sonestedt, E., Hallmans, G., Johansson, I., Fedirko, V., Romieux, I., Ferrari, P., Norat, T., Vergnaud, A. C., Riboli, E., and

Linseisen, J. (2013). Meat consumption and mortality - results from the European Prospective Investigation into Cancer and Nutrition. *BMC Medicine*, 11(1):63.

Rothman, K. J. (1974). Synergy and Antagonism in Cause-Effect Relationships. *American Journal of Epidemiology*, 99(6):385–388.

Rothman, K. J. (1995). Causes. *American Journal of Epidemiology*, 141(2):90–95.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins.

Rothman, K. J., Greenland, S., and Walker, A. M. (1980). Concepts of Interaction. *American Journal of Epidemiology*, 112(4):467–470.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25(1):127–141.

Royston, P., Ambler, G., and Sauerbrei, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28(5):964–974.

Royston, P. and Lambert, P. (2011). *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*.

Saracci, R. (1980). Interaction and Synergism. *American Journal of Epidemiology*, 112(4):465–466.

Saracci, R. (1999). Epidemiology in progress: thoughts, tensions and targets. *International journal of epidemiology*, 28(5):S997.

Shardell, M. D., Alley, D. E., Hicks, G. E., El-Kamary, S. S., Miller, R. R., Semba, R. D., and Ferrucci, L. (2011). Low-serum carotenoid concentrations and carotenoid interactions predict mortality in US adults: the Third National Health and Nutrition Examination Survey. *Nutrition Research*, 31(3):178–189.

Siemiatycki, J. and Thomas, D. C. (1981). Biological Models and Statistical Interactions: an Example from Multistage Carcinogenesis. *International Journal of Epidemiology*, 10(4):383–387.

Sinha, R., Cross, A., Graubard, B., Leitzmann, M., and Schatzkin, A. (2009). Meat intake and mortality: A prospective study of over half a million people. *Archives of Internal Medicine*, 169(6):562–571.

Skrondal, A. (2003). Interaction as Departure from Additivity in Case-Control Studies: A Cautionary Note. *American Journal of Epidemiology*, 158(3):251–258.

Snapinn, S. and Jiang, Q. (2011). On the clinical meaningfulness of a treatment's effect on a time-to-event variable. *Statistics in Medicine*, 30(19):2341–2348.

Stamatakis, K. A. and Punjabi, N. M. (2007). Long sleep duration: A risk to health or a marker of risk? *Sleep medicine reviews*, 11(5):337–339.

Steenland, K. and Deddens, J. A. (2004). A Practical Guide to Dose-Response Analyses and Risk Assessment in Occupational Epidemiology:. *Epidemiology*, 15(1):63–70.

Steffen, L. M., Jacobs, D. R., Stevens, J., Shahar, E., Carithers, T., and Folsom, A. R. (2003). Associations of whole-grain, refined-grain, and fruit and vegetable consumption with risks of all-cause mortality and incident coronary artery disease and ischemic stroke: the Atherosclerosis Risk in Communities (ARIC) Study. *The American Journal of Clinical Nutrition*, 78(3):383–390.

Stranges, S., Dorn, J. M., Shipley, M. J., Kandala, N. B., Trevisan, M., Miller, M. A., Donahue, R. P., Hovey, K. M., Ferrie, J. E., Marmot, M. G., and Cappuccio, F. P. (2008). Correlates of Short and Long Sleep Duration: A Cross-Cultural Comparison Between the United Kingdom and the United States The Whitehall II Study and the Western New York Health Study. *American Journal of Epidemiology*, 168(12):1353–1364.

Takata, Y., Shu, X. O., Gao, Y. T., Li, H., Zhang, X., Gao, J., Cai, H., Yang, G., Xiang, Y.-B., and Zheng, W. (2013). Red Meat and Poultry Intakes and Risk of Total and Cause-Specific Mortality: Results from Cohort Studies of Chinese Adults in Shanghai. *PLoS ONE*, 8(2):e56963.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer New York, New York, NY.

Thiébaut, A. C. M. and Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine*, 23(24):3803–3820.

Thompson, W. D. (1991). Effect modification and the limits of biological inference from epidemiologic data. *Journal of Clinical Epidemiology*, 44(3):221–232.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer,

M., and Wei, L. J. (2014). Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. *Journal of Clinical Oncology*, 32(22):2380–2385.

Uno, H., Wittes, J., Fu, H., Solomon, S. D., Claggett, B., Tian, L., Cai, T., Pfeffer, M. A., Evans, S. R., and Wei, L. J. (2015). Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. *Annals of Internal Medicine*, 163(2):127.

Van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers. *Nature*.

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, T. J. and Knol, M. J. (2014). A Tutorial on Interaction. *Epidemiologic Methods*, 3(1):33–72.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., and Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. *Preventive Medicine*, 45(4):247–251.

Waart, F. D., Schouten, E. G., Stalenhoef, A. F. H., and Kok, F. J. (2001). Serum carotenoids, alpha-tocopherol and mortality risk in a prospective study among Dutch elderly. *International Journal of Epidemiology*, 30(1):136–143.

Wagner, H. M. (1959). Linear programming techniques for regression analysis. *Journal of the American Statistical Association*, 54.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 104(487).

Weed, D. L. (2001). Methods in epidemiology and public health: does practice match theory? *Journal of Epidemiology and Community Health*, 55(2):104–110.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879.

Westreich, D., Cole, S. R., Tien, P. C., Chmiel, J. S., Kingsley, L., Funk, M. J., Anastos, K., and Jacobson, L. P. (2010). Time Scale and Adjusted Survival Curves for Marginal Structural Cox Models. *American Journal of Epidemiology*, page kwp418.

World Health Organization (2002). *Diet, nutrition, and the prevention of chronic diseases: report of a WHO Study Group*, volume 797. World Health Organization.

World Health Organization (2004). Global Strategy on Diet, Physical Activity and Health.

World Health Organization (2009). *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*. World Health Organization.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Yuan, Y. and Yin, G. (2010). Bayesian Quantile Regression for Longitudinal Studies with Nonignorable Missing Data. *Biometrics*, 66(1):105–114.

Yue, Y. R. and Rue, H. a. (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis*, 55(1):84–96.