



Università degli Studi di Milano - Bicocca

FACOLTÀ DI SCIENZE STATISTICHE
Corso di Dottorato di Ricerca in Statistica XXV ciclo

Causal mediation analysis on survival data: an application on the National March Cohort

Discussant:
Alessandra Grotta

Matricola 059195

Tutor:
Prof.Rino Bellocco

Contents

Introduction	1
1 Causal Inference	4
1.1 The concept of cause	4
1.2 Rubin causal model	5
1.3 Causal effect	6
1.4 Direct acyclic graphs	11
2 Mediation Analysis	15
2.1 The concept of mediation	15
2.2 Continuous outcomes	20
2.3 Non continuous outcomes	23
3 Causal Mediation Analysis	28
3.1 Causal framework for mediation analysis	28
3.2 Controlled and natural effects	29
3.3 An illustration: odds ratio decomposition	34
4 Causal Mediation Analysis With Survival Data	38
4.1 Total, direct and indirect effect	38
4.2 Aalen additive hazard model	40
4.3 Cox proportional hazard model	42
4.4 Accelerated time failure model	44

<i>CONTENTS</i>	ii
5 Prostate Cancer	46
5.1 Epidemiology	46
5.2 The role of physical activity	51
5.3 The role of waist-to-hip ratio	54
6 The Swedish Health Care System	57
6.1 General overview	57
6.2 The Swedish National Registries	59
6.3 National Prostate Cancer Register	64
7 Physical Activity and Prostate Cancer	67
7.1 Introduction	67
7.2 Materials	68
7.3 Methods	72
7.4 Results	74
7.5 Discussion	88
Conclusions	93
Bibliography	97
	104

Introduction

Causal inference represents today one important trend of modern research in statistical science. Causal inference is a general term to refer to statistical approaches that share the same underlying philosophy based on counterfactual. The common aim of these techniques grounds on identifying assumptions needed to estimate measures that quantify the strength of causal relations among variables.

Causal inference is an approach that has revealed to be successful in many applied research fields: economists, sociologists and psychologists are increasingly using methods developed in this framework. These techniques have become very popular also in the applied medical research field, for epidemiological and public health studies.

Due to this connection between causal inference theory and medicine field, important epidemiological issues, such that of confounding, have been revised in this new perspective. Often some of the results obtained through causal inference approach have revealed to be equal to ones obtained through traditional statistic methods that have been used during the past decades. In other situations, the causal inference approach was able to show that the commonly used measures did not have any causal meaning and other times these new methods showed the assumptions under which traditional measures could have causal interpretation.

Causal inference had the great merit of having formalized epidemiological concepts that during the years had been developed with different notations in different research fields. In the same way as it was done for confounding, also issue

of moderation and mediation have been reread in this new approach.

In particular, causal inference has allowed a generalization of previous theoretical results present in literature for mediation analysis. Mediation is an important theme in sociological, psychological and medical research field and it arises when we want to test if the effect of a treatment on a certain outcome goes or not through pathways involving one or more other variables, called for this reason mediators.

When dealing with mediation, most of the applied works were referring to the influential paper of Baron and Kenny of 1986. However, the theory that the authors developed is mainly focused on situations in which the outcome is linear and in which there is no interaction between the treatment and the mediator variable. An extension of this theory has been possible thanks to the formalization through counterfactual notation of controlled and natural effects. Work of authors like Pearl, Rubin and VanderWeele has allowed researchers to handle situations in which treatment-mediator interaction is present or in which there are dichotomous and survival outcomes.

The aim of this thesis is to use the recent development of this theory to answer a relevant clinical question in the field of prostate cancer epidemiology. Prostate cancer is one of the most common male cancer in USA and Europe and its incidence has increased in all developed countries in the last decades. Prostate cancer etiology has not been completely understood and the only well known established factors are increasing age, ethnicity and heredity. These factors are unmodifiable risk one, whose knowledge can help the physician to detect subjects at higher risk and make earlier diagnoses through screening. However, it would be important to identify modifiable life style factors that could be changed in order to prevent prostate cancer onset.

A big debate has been going on in literature about the possible effect of physical activity in preventing this tumor. Physical activity has indeed been found to have positive effect on many cancers and there are plausible biological mechanisms that could be involved also in prostate carcinogenesis.

It is well known that physical activity can affect hormonal metabolism. Physical

activity decreases levels of serum testosterone and sex hormone binding globulin, that in turn can affect the risk of prostate cancer development. However, studies assessing the relationship of physical activity with prostate cancer risk are inconclusive.

Similar pathways have been suggested also to explain a possible role of waist-to-hip ratio control in the tumor prevention. Abdominal adiposity affects levels of testosterone and sex hormone binding globulin, increases levels of insulin and leptin that, in turn, modifies hormonal metabolism. Results available in literature are however not consistent.

Our aim is to understand if physical activity effect on prostate cancer goes through waist-to-hip ratio reduction. Consequently, the purposes of this investigation are to analyze the association of physical activity and waist-to-hip ratio on prostate cancer and to analyze if waist-to-hip ratio can mediate the effect of physical activity on prostate cancer incidence and mortality. Separated analyses will be performed for total, low and high-risk prostate cancer. We will study these associations among a cohort of 13.000 Swedish residents who were followed up from 1997 to 2010.

Chapter 1

Causal Inference

1.1 The concept of cause

Before introducing causal inference, it is interesting to review the old ideas of cause, a concept on which philosophers have been discussing for millennia. In the Western philosophical tradition, Aristotle was one of the first who began a discussion on this topic. In his theory on causation he listed four causes of a thing in his physics: the material cause (that out of which the thing is made), the formal cause (that into which the thing is made), the efficient cause (that which makes the thing), and the final cause (that for which the thing is made). His notion of efficient cause is the one that is close to the usual modern definition of cause. However, one of the biggest limitation of his theory was that he defined cause without reference to any effect of the cause.

The concept of effect as a consequence of a cause was introduced only later, with Locke in the 17th century. Locke introduced his own definitions on causality and basic to his approach to the concept of causation was the idea of power. He held the Aristotelian belief that causes are substantial powers put to work and wrote “... *power being the source from whence all action proceeds, the substances wherein these powers are, when they exert this power into act, are called causes; and the substances which thereupon are produced are called effects* ...”.

A different approach to causation was proposed by Hume in the 18th century,

who introduced three basic criteria for causation: *i*) spatial/temporal contiguity, *ii*) temporal succession, and *iii*) constant conjunction. Thus for Hume, in order to show A causes B , it is necessary that *i*) A and B be contiguous in space and time, *ii*) A precede B in time, and *iii*) A and B occur (or do not occur) together.

Mill in the 19th century provided some ideas regarding how to discover causation in practice. He described four methods: *i*) the method of concomitant variation: if Y varies as A varies, A might be a cause of the change in Y ; *ii*) the method of difference: the difference between Y when A happens and when B happens indicates the cause; *iii*) the method of residuals: the effect of B on Y can be observed by taking the difference between Y when A and B both happen and that when only A happens; and *iv*) the method of agreement: if Y does not change regardless of A or B happening, neither A nor B cause change in Y .

Reasoning on the concept of cause is still a topic in contemporary philosophy and there has been an increasingly connection between philosophical and other research fields. The concept of cause has interested not only philosophers but also scientists whose aims is often that of identify relations among variable that have a causal interpretation.

1.2 Rubin causal model

In the field of causal inference there are today two main competing approaches, the counterfactual perspective and noncounterfactual perspective. One particular model based on the counterfactual perspective has been developed by Rubin over the last 35 years [1] and today it is known as the potential outcomes framework or the Rubin causal model.

This model has its roots in the philosophical work of Hume and it was developed in the field of randomized experiment by Neyman [2] and Fisher [3] and only successively it has been extended to observational context by Rubin [4].

Potential outcome definition

Suppose the case in which there is a binary treatment variable A that for subject i takes value equal to 1 if the individual receives the treatment and that takes value equal to 0 otherwise. Let X represents a set of variables that take their values before the treatment assignment or, more generally, that are not affected by the received treatment. Let Y be the response variable, that is a variable of scientific interest and suppose we are interested in how it is causally related to variable A . For each individual i , it is possible to define $Y_i(0)$ and $Y_i(1)$ as the values of response variable if subject receives respectively treatment 1 or 0.

In Rubin's model only one of the two treatments can be applied, and only the response to that treatment can be observed. Once treatment 1 is applied to subject i , $Y_i(0)$ becomes a counterfactual since treatment 0 was not applied. As a result, $Y_i(0)$ cannot be observed. Similarly, if treatment 0 is applied to unit i , $Y_i(0)$ is observed and $Y_i(1)$ is the counterfactual. In other words, one of the potential responses would become the actual response and the other would become the counterfactual.

Most analyses, make the stable unit-treatment value assumption (SUTVA) [5], meaning that the response of a subject i to treatment A is not affected by what treatments other subjects receive. This means that if unit i receives the same treatment a in two different assignments then, under SUTVA, $Y_i(a)$ would be the same value regardless of how different are the treatment assignments for other units. But it implies also that there are no hidden versions of treatments: no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0.

1.3 Causal effect

We say that the treatment A has a causal effect on outcome Y for subject i if $Y_i(1) \neq Y_i(0)$. The causal effect of treatment A on variable Y for subject i can then be quantified by the following algebraic difference:

$$\tau_i = Y_i(1) - Y_i(0) \quad (1.1)$$

Other types of expressions involving comparisons between two potential outcomes, such as the ratio between them, are meaningful measures of causal effects. Individual causal effects, however, cannot be estimated because of the so called “*Fundamental Problem of Causal Inference*”. The fundamental problem of causal inference states that it is not possible to *observe* both $Y_i(0)$ and $Y_i(1)$, thus it is not possible to observe the treatment effect for subject i . In fact, for subject i , only the counterfactual outcome that corresponds to the treatment value that the subject actually was assigned becomes a factual. Each potential outcome is observable, but we can never observe all of them. It is not possible then to identify individual causal effects because of missing data.

Whereas it is not possible to estimate individual effects, it is possible to define and estimate measures that quantify average causal effects. These measures are not able to tell us anything about the effect of treatment A on subject i , but can define and estimate average measures of causal effect on a defined population. More in detail, we say that a treatment A has a causal effect on the outcome Y on a certain population if the following statement hold:

$$E[Y_0] \neq E[Y_1] \quad (1.2)$$

Measures involving comparisons between these two quantities can be then be defined to quantify the causal effect on population level. Without any loss of generality, we can assume, for example, Y to be a dichotomous variable and we can define risk difference (RD), relative risk (RR) and odds ratio (OR) in the following way:

$$RD = E[Y_1] - E[Y_0] \quad (1.3)$$

$$RR = E[Y_1] - E[Y_0] \quad (1.4)$$

$$OR = \frac{E[Y_1]}{1 - E[Y_1]} / \frac{E[Y_0]}{1 - E[Y_0]} \quad (1.5)$$

In literature, different estimation methods have been proposed. Very popular are methods based on propensity scores [6]. A complete overview can be found in D’Agostino [7].

The assignment mechanism

Let A_i indicate the assignment for unit i and define vector A as the vector for all units assignments, that is $A = (A_1, \dots, A_i, \dots, A_N)^T$. The assignment mechanism gives the conditional probability of each vector of assignments given the vector of covariates and vectors of potential outcomes:

$$Pr(A \mid X, Y(1), Y(0)) \quad (1.6)$$

An example of an assignment mechanism is a completely randomized experiment with $N = 2n$ units. In this situation we have:

$$Pr(A \mid X, Y(1), Y(0)) = \begin{cases} 1/C_n^N & \text{if } \sum W_i = n, \\ 0 & \text{otherwise .} \end{cases}$$

An “unconfounded assignment mechanism” is free of dependence on either $Y(0)$ or $Y(1)$:

$$Pr(A \mid X, Y(1), Y(0)) = Pr(A \mid X) \quad (1.7)$$

With an unconfounded assignment mechanism, at each set of values of X_i that

has a distinct probability of $A_i = 1$, there is effectively a randomized experiment. That is, if X_i indicates sex, with males having probability 0.2 of receiving the active treatment and females probability 0.5, then essentially one randomized experiment has been done for males and another for females.

The assignment mechanism is “probabilistic” if each unit has a positive probability of receiving either treatment, that is:

$$0 < Pr(A | X, Y(1), Y(0)) < 1 \quad (1.8)$$

A strongly ignorable treatment assignment is an assignment both unconfounded and probabilistic.

The assignment mechanism is fundamental to causal inference because it tells us how we got to see what we saw. Causal inference is basically a missing data problem because at least half of the potential outcomes are not observed, and so missing. Without understanding the process that creates missing data, we have no hope of inferring anything about them. Without a model for how treatments are assigned to individuals, formal causal inference, at least using probabilistic statements, is impossible. This does not mean that we need to know the assignment mechanism, but rather that without positing one, we cannot make any statistical claims about causal effects, such as unbiasedness of estimates, confidence coverage of intervals for effects, levels of tests of significance, or coverage of Bayesian posterior intervals.

Randomization is an assignment mechanism that allows particularly straightforward estimation of causal effects. Therefore, simple randomized experiments form the basis for inference for causal effects in more complicated situations, such as when assignment probabilities depend on covariates or when there is non-compliance with the assigned treatment. An unconfounded assignment mecha-

nism, which essentially is a set of randomized experiments, forms the basis for the analysis of an observational non randomized study by using the randomized experiment as a template.

Estimating treatment effects

Relying only on a model for the assignment mechanism, we can make progress on statistical inference for causal effects, even in observational studies, using, for example, propensity scores [6] and all the related methods that were developed after the explosion of recent literature on propensity scores in applied journals.

Propensity score methods were first introduced by Rosenbaum and Rubin [6]. Their success is due to the fact that there are many situations in which they have been used to uncover answers similar to those in randomized experiments. Propensity score methods have been used in all research fields, including the economical, social and medical science ones.

A key idea with epidemiological data is to try to estimate the assignment mechanism assuming it is unconfounded. That is, try to estimate the “propensity score”, the probability of being assigned the treatment as a function of all covariates (but not the potential outcomes, or their partially observed values). Suppose to have a dichotomous treatment, then the propensity score $e(X)$ is then defined as:

$$e(X) = P(A = 1 | X) \tag{1.9}$$

Having estimates of the propensity scores for all the subjects, it is possible to form groups of treated or control units with approximately equal values of the propensity scores. For example, one group could consist of all units with estimated propensity scores between 0.2 and 0.3 (such a group will have relatively more control than treated because the probability of being treated is less than

0.5), but all units in that group will be estimated to be approximately equally likely to be treated. Thus, in this group, assuming the assignment mechanism was unconfounded given the observed covariates, a randomized experiment was essentially done. An implication of this logic is that within a group of units with nearly constant propensity score, the multivariate distribution of X_i will be as in a randomized experiment, that is, approximately the same for treated and control units. Thus, the following relation holds:

$$A \perp X \mid e(X) \tag{1.10}$$

This proposition is very important practically because it is easily checked from observed data on the covariates.

The problem of causal inference becomes that of finding the minimal set of variables conditional on which the treatment assignment can be defined unconfounded.

1.4 Direct acyclic graphs

When reasoning on causal relations among a set of variables, some set of assumptions need to be verified. Sometimes the assumptions of an analysis model are clearly correct. However, most epidemiological research is plagued by uncertainty about assumptions: in these cases it is important to recognize and explicate fully the analysis model.

A serious drawback of common statistical models is that they embody many parametric assumptions that are not known to be correct and may be incorrect. Another drawback of common statistical models is that they cannot capture all types of assumptions.

Causal diagrams are graphical models for causal relations that can serve a role complementary to conventional models; such diagrams do not incorporate the strong parametric assumptions of conventional models; instead, they display assumptions about the web of causation that are not captured by conventional

models.

For example, theoretical development on directed acyclic graphs (DAGs) has been shown to be useful in the causal inference research field [8]. Their use has indeed revealed fundamental to understand conditioning on which set of covariates X a treatment assignment A can be defined unconfounded.

A DAG is a mathematical tool composed of nodes and directed edges between nodes whose aim is to represent relations between interconnected objects. This graph is directed because connections between two nodes are represented by arrows, and acyclic, since following the arrows it is not possible to return to the starting node.

Use of DAGs has become very popular in the epidemiological field to formalize relations between observed and unobserved variables involved into a study. In this situation, variables are represented by nodes.

Conditional independence

When dealing with DAGs, we define a *path* to be a route between two variables passing through the arrows following or not their directions.

The ancestors of a variable V are all other variables which affects V either directly or indirectly.

The descendants of a variable V are all other variables affected by V , either directly or indirectly.

If all paths between two variables are blocked by conditioning on a set of variables, then the two variables are said to be “conditionally independent” given the set. On the contrary, if one open path exists, then the variables are conditionally associated. To understand if two variables are independent given a set of other variables, Pearl [9] developed the following algorithm: a path can be blocked or open according to two very simple rules:

- a path is blocked if it contains a chain $\leftarrow L \rightarrow$ or a $\leftarrow L \rightarrow$ and we have conditioned on the middle variable
- path is blocked if it contains an inverted fork $\rightarrow L \leftarrow$ and we have not

conditioned on the middle variable or on any of its descendants (the middle variable in an inverted fork is called *collider*)

In all the other cases, a path is open. Statistical associations among variables can be determined either by blocked and unblocked paths.

To test assumptions of conditional independence Pearl [10] developed a graphical criterion known as d-separation.

D-separation

A set S of nodes is said to block a path if either *i*) the path contains at least one arrow-emitting node that is in S , or *ii*) the path contains at least one collision node that is outside S and has no descendant in S . If S blocks all paths from X to Y , it is said to *d-separate* X and Y and then, X and Y are independent given S , that is $X \perp Y \mid S$.

Causal direct acyclic graph

When using DAGs for causal inference purposes, it is necessary to define the concept of causal DAG. In a causal DAG directions of arrows have a precise meaning since they give the direction of the possible causal relation from one variable to another. Presence of arrows encodes assumptions: if an arrow from L to A is absent then we believe that variable L does not affect variable A .

Thus we could say that causal assumptions are encoded not in the links but, rather, in the missing links. An arrow indicates only the possibility of causal connection, the strength of which remains to be determined from data.

When dealing with causal DAGs we assume that all common causes of pairs of variables on the graph are themselves on the graph and in which we can assume that any variable is a cause of its descendants. In doing that, causal DAGs encode all the causal determinants of statistical associations.

DAGs thus help the researcher to distinguish between causal relationship among variables and statistical associations that can arise, for example, when two variables share a common causes.

Besides the concept of causal DAG, it is possible to define the concept of minimal causal DAG. We say that a causal DAG is the minimal one if the graphical representation involves only sufficient causes.

Minimal causal DAGs are strictly related to Rothmans sufficient-component causes framework [11]. Within this framework, causation is conceptualized as a series of different causal mechanisms, each sufficient to bring about the outcome. These causal mechanisms are called by Rothman “sufficient causes”. Rothman conceived of them as minimal sets of actions, events, or states of nature which together initiated a process resulting in the outcome. For a particular outcome there would likely be many different sufficient causes, that is, many different causal mechanisms by which the outcome could come about. Each sufficient cause involved various component causes. Whenever all components of a particular sufficient cause were present, the outcome would inevitably occur; within every sufficient cause, each component would be necessary for that sufficient cause to lead to the outcome.

It is sometimes an advantage to reduce a redundant set of sufficient causes to a nonredundant set of minimal sufficient causes. This happens because if we allow in the graph sufficient causes that are not minimally sufficient or redundant sufficient causes or redundant minimal sufficient causes it is less straightforward to highlight the conditional independence relations implied by the structure of the causal directed acyclic graph.

Chapter 2

Mediation Analysis

2.1 The concept of mediation

Most of the research in the social, economical and medical fields focuses on the relation between two variables. For example, we could be interested in the causal effect of a treatment variable on an outcome. Let A be the treatment variable and Y the outcome variable. The setting under study is diagrammed in Figure 2.1.

This situation could be complicated by the presence of a third variable Z . Different are the scenarios that could arise by introducing this new covariate. In this paragraph we will present three important scenarios in which the third variable under study leads to the situations of confounding, interaction and mediation.



Figure 2.1: DAG representing causal relationship between a treatment variable A and outcome Y

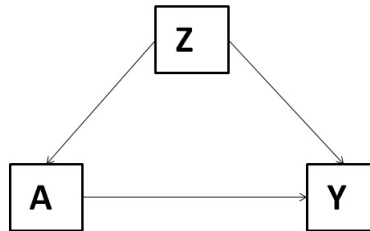


Figure 2.2: DAG representing causal relationship between a treatment variable A and outcome Y , confounded by variable Z

Confounding

The first setting that we could imagine is the one in which variable Z affects both variable X and variable Y : in this case, it would behave like a *confounder* and not taking it into account when analyzing data would lead to biased estimates of the causal relation between the treatment and the outcome. The setting would be that shown in Figure 2.2.

Variable Z would act as a confounder since there is an open path from variable A to variable Y . For this reason we must control for this variable when analyzing data. In contrast with the classical definition of counfounder, we indeed refer to Hernan [12] who writes: “*One possibility would be to define a confounder as a variable that, possibly in conjunction with other variables, can be used to ensure all otherwise open backdoor paths between treatment and outcome are blocked*”.

Moderation

The second scenario we could hypothesize is that in which Z acts as an *moderator*, that is the relation between the treatment and the outcome changes as a function of variable Z .

When dealing with moderator variables, the relation between the treatment and the outcome can change not only for what concerns the strength of the relation but also for what regards the direction.

The issue of moderation, also known as interaction, is a central one in both

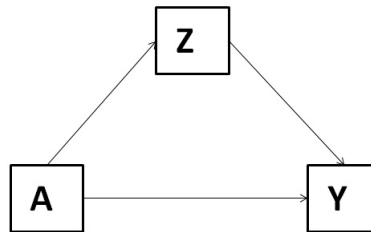


Figure 2.3: DAG representing causal relationship between a treatment variable A and outcome Y , mediated by variable Z

social and epidemiological research and has been well developed in scientific literature. A landmark paper was published in 1986 by Baron and Kenny [13]: in this article they analyzed possible possible scenarios in which moderation arises and they considered specific analysis procedures for appropriately measuring and testing hypotheses of moderation.

More in detail, they analyzed situation in which *a*) both the moderator and the outcome are dichotomous *b*) the moderator is continuous and the outcome is dichotomous *c*) the moderator is dichotomous and the outcome is continuous *d*) both the moderator and the outcome are continuous. Possible settings in which moderation arises are presented by Hernan [12].

Mediation

A third setting we could create is the one in which variable Z is on the causal sequence between A and Y . In this case variable Z acts as a *mediator*.

The graphical representation is diagrammed in Figure 2.3. In this case we have two paths going from treatment to outcome variable: the first one that is direct and the second one that goes through a third variable M . For this reason the latter is called the indirect, or mediated, path.

When the direct path is almost null, then we have evidence for a single and strong mediator, otherwise multiple factors are probably acting as mediators.

Statistical methods to address mediation issue have become more popular after the work of Judy and Kenny [14] and of Baron and Kenny [13]. Their work was

in the psychological field but today mediation analysis plays a critical role also in social, economical and medical sciences since it helps understanding processes that underlie the path between two variables. Thousands of hypothesis relating to mediation issues are presented in literature and there is a growing demand for methods that can address in an appropriate way complex questions.

Moderation and Mediation

In many situation, we could be interested in assessing if a mediation effect is constant or changes depending on subgroups: for example, a certain variable Z could mediate the relationship between an exposure A and an outcome Y only for men but not women. We could also imagine that the intensity of a mediated effect may depend on the value of a moderator Z in a linear way.

To test these hypotheses, researchers need appropriate statistical models. Even if models assessing interaction have been always treated apart from models dealing with mediation, this separation does not reflect real possible observational situations, in which interaction and mediation can be found and combined together.

This issue has been well analyzed by Baron and Kenny [13] and between *moderated mediation* and *mediated moderation*.

Moderated mediation

Moderated mediation is a term coined by James and Brett in 1984 [15]. The moderator mediation setting is a simple scenario that deals with both mediation and moderation issues. In this model, there is a mediation variable between the treatment and the outcome variables, but the mediation intensity depends on the values taken by a moderator variable. Thus, for subgroups of subjects we have different mediation patterns.

Five different possible ways in which a mediated effect can be function of a moderator variable have been proposed:

- the independent variable A functions as a moderator of the mediator-outcome path

- some fourth variable W_1 affects the treatment-mediator path
- W affects the mediator-outcome path
- W affects the treatment-mediator path whereas another variable W_2 affects the mediator-outcome path
- W affects both the treatment-mediator and the mediator-outcome path

Mathematical details and graphical representations of the depicted scenarios can be found in the work of Preacher [16].

Mediated moderation

Moderation may also involve a mediator variable [13, 15, 17]. In this case, the interaction effect of the independent and moderator variables on the dependent variable is transmitted through the mediator variable.

A prerequisite of mediated moderation is the occurrence of overall moderation between the independent and dependent variables [13]. The effect of the independent variable on the dependent variable must depend on the moderator variable. There are at least three different types of mediated moderation: between the independent and mediator variables, between the mediator and dependent variables, or both [17].

Mediated moderation can be used to explain the causal relationship between four variables. For instance, Scheufele [18] proposes that interpersonal discussion of politics serves as a moderator in the relationship between hard news use and political participation, and this moderation is further mediated by political knowledge. For people with high levels of self-reported political discussion, hard news use leads to increased political knowledge, which enhances political participation. On the other hand, for people with low levels of self-reported political discussion, hard news use has no significant influence on political knowledge or participation.

2.2 Continuous outcomes

The most famous approach to mediation analysis is the one developed by Baron and Kenny [13] for continuous outcomes and it is based on linear structural equation modeling. To present this approach to mediation, we first have to define the following set of equation:

$$Y = \alpha_1 + cA + \epsilon_1 \quad (2.1)$$

$$Y = \alpha_2 + c'A + bM + \epsilon_2 \quad (2.2)$$

$$M = \beta + aA + \epsilon_3 \quad (2.3)$$

where Y is a continuous outcome variable, A is the treatment variable, M is the mediator.

In Equation 2.1 the outcome variable is regressed on the treatment variable, in Equation 2.2 the outcome variable is regressed on both the treatment and the mediating variables and in Equation 2.3 the mediating variable is regressed on the treatment variable. Equations can be also modified to account for nonlinear effects and for interactions between the treatment and the mediator. To assess the presence of mediation, Baron and Kenny developed a four steps approach. After estimating the coefficient of the models proposed, it is necessary to find a) a significant relation between the treatment and the outcome b) a significant relation between the mediator and the outcome c) a significant relation between the mediator and both the outcome and treatment variable. Moreover, the relation between the treatment and the outcome has to be larger than the same relation in the model containing also the mediator variable.

Difference and product method

To quantify the mediated effect, two different approaches are possible, the difference and the product method.

The difference method consists in estimating Equations 2.2 and 2.3 and calculate the indirect effect as $c - c'$. The rationale behind this method is that we

quantify the reduction of the relation between the treatment and the outcome variable after introducing in the model the mediator variable. This method has become very popular in the epidemiological field. The product method consists in estimating Equations 2.2 and 2.3 and calculate the indirect effect as $\hat{a}\hat{b}$. The reason for doing that is that mediation should depend on both how much the exposure is able to influence the mediator as well on how much the mediator is able to influence the outcome.

It has been shown that, when the outcome is continuous, $\hat{a}\hat{b}$ and $\hat{c} - \hat{c}'$ quantities are equal if the mediation regression equations are estimated through ordinary least squares regression and if the residual variance follows a normal distribution [19]. This result is no longer true if we use logit or probit models or if we are in the context of survival analysis.

Significance testing and confidence intervals

To compute the significance level, both the product or the difference quantities can be divided by their standard error and the ratio is then compared to the significance level of a standard normal distribution.

Many formulas have been proposed in literature for the standard error of the mediated effect. When using the difference method standard error can be obtained as:

$$\sigma_{\hat{c}-\hat{c}'} = \sqrt{\sigma_{\hat{c}}^2 + \sigma_{\hat{c}'}^2 - 2r\sigma_{\hat{c}}\sigma_{\hat{c}'}} \quad (2.4)$$

Most of the literature has focused on the standard error for the product quantity. Assuming that the product follows a normal distribution, then the exact standard error is the following one:

$$\sigma_{\hat{a}\hat{b}} = \sqrt{\sigma_b^2 \hat{a}^2 + \sigma_a^2 \hat{b}^2 + \sigma_a^2 \sigma_b^2} \quad (2.5)$$

This formula was proposed by Arolan [20] but other authors found it through

delta method and a second order Taylor series. When the term is sufficiently small it can be omitted by formula and we obtain an asymptotic result that is the same the Sobel [21] found through a first order delta method, that is:

$$\sigma_{\hat{a}\hat{b}} = \sqrt{\sigma_a^2 \hat{b}^2 + \sigma_b^2 \hat{a}^2} \quad (2.6)$$

With the same consideration, we can build confidence intervals for the mediated effect, that is:

$$CI_{1-\alpha} : \hat{a}\hat{b} \pm z_{\alpha/2} \sigma_{\hat{a}\hat{b}} \quad (2.7)$$

These procedures are very simple to apply but the problem associated with them is that both hypothesis tests and confidence intervals are based on the assumption that the distribution of the product is a standard normal. Many statistical books have discussed the issue of the variance of the product of two independent normal variables and it is known that the normal distribution for the product is valid only in particular cases or if the sample is large enough, otherwise the distribution will be kurtotic and positively skewed. For example, the product of two standard normal variables has an excess kurtosis of six compared to the kurtosis of a normal distribution. For this reason, methods based on normal approximation are often inaccurate and it was shown that this results in not accurate type I error rates and statistical power. If then accurate results are needed other methods can be used. Among them one of the most popular is bootstrapping or re-sampling. Bootstrapping requires to draw N units with replacement from the original sample of size N . Thus the empirical distribution obtained through sampling can be used to get asymmetric confidence intervals that do not rely on the normality assumption.

2.3 Non continuous outcomes

Under the assumptions before stated the results derived using difference and product method coincide. As mentioned earlier, however, these results are no more valid when using model that have dichotomous or time to event outcomes variables.

Dichotomous outcomes

To present the approach to mediation analysis when dealing with dichotomous outcomes, we present the example shown in Karlson and Anders [22]. Suppose the outcome of interest is a dichotomous variable Y as, for example, passing the admission university test. Let A be a continuous or categorical variable measuring social origin and let Z be a measure of academic performance. In order to perform mediation analysis, we can start thinking of variable Y as a dichotomous version of a continuous latent variable Y^* .

$$Y^* = \alpha + \beta A + \gamma M + v \quad (2.8)$$

where v is a gaussian error with mean equal to 0 and variance equal to σ_v^2 . In the same way, suppose the following model for the mediator to hold:

$$M = \mu + \theta A + \nu \quad (2.9)$$

where ν is a gaussian error with mean equal to 0 and variance equal to σ_ν^2 . We cannot estimate parameters of model 2.8 since the outcome variable is latent. However, we observe Y that is the result from the following process:

$$Y = \begin{cases} 1 & \text{if } Y^* > \tau, \\ 0 & \text{if otherwise.} \end{cases} \quad (2.10)$$

Usually, τ is set equal to 0. Error v can be rewritten as $v = \sigma_e \omega$ where ω is a

logistic random variable with mean equal to 0 and variance equal to $\pi^2/3$ and σ_e is a scale parameter $\sigma_v = \sigma_e(\pi/\sqrt{3})$. The probability of Y in function of A can be expressed as:

$$\begin{aligned}
 Pr(Y = 1) &= Pr(Y^* > 0) \\
 &= Pr\left(u > -\left[\frac{\alpha}{\sigma_e} + \frac{\beta}{\sigma_e}A + \frac{\gamma}{\sigma_e}M\right]\right) \\
 &= F\left(\frac{\alpha}{\sigma_e} + \frac{\beta}{\sigma_e}A + \frac{\gamma}{\sigma_e}M\right) \\
 &= \frac{\exp\left(\frac{\alpha}{\sigma_e} + \frac{\beta}{\sigma_e}A + \frac{\gamma}{\sigma_e}M\right)}{1 + \exp\left(\frac{\alpha}{\sigma_e} + \frac{\beta}{\sigma_e}A + \frac{\gamma}{\sigma_e}M\right)}
 \end{aligned} \tag{2.12}$$

where $F(\cdot)$ is the cumulative logit function. If we take the logarithm of the odds then we get:

$$\text{logit}(Pr(Y = 1)) = a + bA + cM = \frac{\alpha}{\sigma_e} + \frac{\beta}{\sigma_e}A + \frac{\gamma}{\sigma_e}M \tag{2.13}$$

The parameters of the previous model are the same of the underlying linear model except for the fact that they are divided by the residual deviation of that model multiplied by a constant. Then, when using a logit model, we can identify the underlying coefficients relative to a scale, function of the residual standard deviation of the underlying linear model. Then, we have:

$$b = \frac{\beta}{\sigma_e}; \quad c = \frac{\gamma}{\sigma_e}$$

This fact, however, does not prevent us from decomposing a total effect in a part due the mediator variable and to an other that is not attributable to it.

All effects are measured on the same scale, making them directly comparable. Following Karlson, we can decompose a total effect into a direct and indirect as follows:

$$\begin{aligned} \text{Direct :} \quad & b = \frac{\beta}{\sigma_e}; \quad c = \frac{\gamma}{\sigma_e} \\ \text{Indirect :} \quad & c \cdot \theta = \frac{\gamma}{\sigma_e} = \frac{\theta\gamma}{\sigma_e} \\ \text{Total :} \quad & \frac{\delta}{\sigma_e} = \frac{\beta}{\sigma_e} + \frac{\theta\gamma}{\sigma_e} = \frac{\beta + \theta\gamma}{\sigma_e} \end{aligned}$$

Since these effects are expressed on the same scale, it is possible to compute ratio that tell us about the relative magnitude of direct and indirect effects. This can be done as follows:

$$\begin{aligned} \text{Direct} &= \frac{b}{b + c + \theta} = \frac{\beta/\sigma_e}{\delta/\sigma_e} = \frac{\beta}{\beta + \gamma\theta} \\ \text{Indirect} &= \frac{c\theta}{b + c\theta} = \frac{\gamma\theta/\sigma_e}{\delta/\sigma_e} = \frac{\gamma\theta}{\beta + \gamma\theta} \end{aligned}$$

This method can be extended to control for confounding variable as shown in Karlson [22].

Survival outcomes

For what regards survival outcomes, Tein and MacKinnon [23] proposed a simulation study in which they examined if the same results present in literature for a continuous outcome, apply also when the outcome is a time to event variable. In this study the authors generated an independent treatment variable A and a mediating variables according to model 2.2. Then they generated a time to event variable T through the Weibull model:

$$\log(T_i) = u + c'A_i + \beta M_i + \sigma W \quad (2.14)$$

where the residual part σW is the product of an extreme value distribution W and σ that is a scale parameter on which the shape of the hazard rate depends. After generating data, the authors implemented both a log-survival time model:

$$\log(T_i) = \beta_0 + \beta_{i1}X_1 + \dots + \beta_{ik}X_k + \sigma\epsilon_i \quad (2.15)$$

and a Cox proportional hazard model:

$$\log\lambda(t) = \log\lambda_0(t) + \beta'_{i1}X_1 + \dots + \beta'_{ik}X_k \quad (2.16)$$

It has been shown that for Weibull model the relation $\beta'_j = \beta_j/\sigma$ holds. When using a log-survival time model, the mediation component is expressed in terms of percent of increase (or decrease) of the expected survival time, while when using Cox model the mediation component is expressed in terms of percentage change in hazard rate. Parameters of the log-survival time model were estimated through maximum likelihood method while for Cox model maximum partial likelihood was used. Data were generated without censoring, with no ties of the events and with time invariant mediating variable.

Both the difference and the product method previously described for continuous outcomes were applied for the log-survival time model and for log-hazard model. Results show that, under the previous assumption, difference and product method gave almost identical results when using log-survival time model. Moreover, the estimated $\hat{c} - \hat{c}'$ and $\hat{a}\hat{b}$ are similar when simulation sample size changes. However when using log-hazard model these two methods lead to different result and $\hat{a}\hat{b}$ quantity changed in absolute value and decreased when the simulation sample size grew.

When analyzing confidence intervals and significance tests through the ratio $\hat{a}\hat{b}/\hat{\sigma}_{ab}$ results were almost comparable when using log-survival time model and log-hazard model. Summarizing, the authors have proposed a scenario in which

the theory that was developed by Baron and Kenny for continuous outcomes cannot be applied in general to survival models and that further research is necessary to understand mediation analysis when dealing with survival outcomes.

Chapter 3

Causal Mediation Analysis

3.1 Causal framework for mediation analysis

The concept of mediation is strictly related to the issues of causal inference. When dealing with mediation we are indeed interested in understanding to which extent the effect of a variable on an outcome is mediated by other variables, that is we aim to quantify the causal effect of a variable on an intermediate one and, in turn, the causal effect of the intermediate one on the outcome.

The aim of this chapter is to formalize mediation analysis in terms of potential outcome framework and to understand which are the assumptions required for the identification of *causal mediation effects*.

We present the approach based on the concepts of controlled and natural effects proposed by Robins and Greenland [24] and we show the extent to which this approach is related to the principal stratification framework proposed by Frangakis and Rubin [25].

The counterfactual framework and notation can be extended to define causal mediation effects. Let M be a variable affected by A and that lies on the causal pathway from A to Y . For example, as done by Imai et al. [26] in the JOBS II study, we could be interested in studying if a job training program has an effect on depression levels and in testing if this effect is mediated by the improved self-confidence level of the worker in searching for a job or in completing the

application request for the employment.

Let M_i be the observed level for subject i of job search efficacy: since this value could be affected by the participation to the training program, we will have to define two different potential values, that is $M_i(1)$ and $M_i(0)$. As in the classical causal inference context, we will observe only one of the two potential values, the one corresponding to the real treatment received.

The concept of potential outcome has now to be redefined. In the classical causal inference framework, potential outcomes are function only of the treatment, but in the causal mediation analysis one potential outcomes depend not only on the treatment variable but also on the mediator one. The notation $Y_i(a, m)$ then is used for the potential outcome of subject i when jointly $A = a$ and $M = m$.

Before proceeding we need the consistency and composition assumptions. Consistency assumption states that potential outcomes $Y_i(a)$ and $M_i(a)$ are respectively equal to Y_i and M_i observed on subject i where treatment A is equal to a and mediator M is equal to m . Composition assumption states that $Y_i(a) = Y_i(a, M(a))$.

As in Rosembaum and Rubin, moreover, we assume no between units interference, that is we assume that the treatment status of a unit does not affect the potential mediator values of an other one and that the mediator and treatment values of a unit does not affect the potential outcomes of the others.

The terminology introduced so far allows us to present the two main approaches at the problem of mediation that have been developed in the causal inference framework.

3.2 Controlled and natural effects

The terminology of controlled direct effect and natural direct and indirect effect was introduced by Pearl [27], building on concepts proposed by Robins and Greenland [24].

The controlled direct effect of treatment A on outcome Y is obtained setting the mediator M to level m . Suppose that we are comparing treatment levels $A = a$ and $A = a^*$, then we say that there is a non-null controlled effect for subject i if $Y_i(a, m) \neq Y_i(a^*, m)$ and we can quantify the controlled effect, for example, as $Y_i(a, m) - Y_i(a^*, m)$. The term “direct” refers to the fact that the effect under study quantifies only the direct link between the treatment and the outcome and it is not mediated by other covariates.

If we want to decompose an effect into direct and indirect components, then controlled direct effects are not useful: indeed it is impossible to fix certain variables values in order to measure the intensity of only the pathways that circumvent the direct one between the treatment and the outcome. However, if we are in the situation in which there is no interaction between variables A and M then we can subtract from the total effect the controlled direct effect and obtain a quantity that can be interpreted as an indirect effect [24].

For this reason, if we are interested in etiology studies, natural direct and indirect effects are more useful. We say that there is a non-null natural direct effect for subject i if $Y_i(a, m_{a^*}) \neq Y_i(a^*, m_{a^*})$. Natural direct effect can be then been quantified, for example, as $Y_i(a, m_{a^*}) - Y_i(a^*, m_{a^*})$: the mediator variable takes the value that it would have taken under reference treatment level a^* . A non-null natural indirect effect exists for subject i if $Y_i(a, m_a) \neq Y_i(a, m_{a^*})$ and it can be quantified, for example, as $Y_i(a, m_a) - Y_i(a, m_{a^*})$ and it measures the effect of the mediator on the outcome, when treatment variable A takes value equal to a .

Natural direct and indirect effect always sum up to a total effect, also when interactions or non-linearities are present. Thus, if we suppose to have both a dichotomous treatment and mediator variables A and M , than the following expression always holds:

$$\begin{aligned}
Y(1) - Y(0) &= Y(1, m_1) - Y(0, m_0) \\
&= (Y(1, m_0) - Y(0, m_0)) + (Y(1, m_1) - Y(1, m_0)) \quad (3.1)
\end{aligned}$$

This decomposition, however, is not the only possible one: in fact, we could have also decomposed the total effect into a direct and indirect ones as follows:

$$\begin{aligned}
Y(1) - Y(0) &= Y(1, m_1) - Y(0, m_0) \\
&= (Y(1, m_1) - Y(0, m_1)) + (Y(0, m_1) - Y(0, m_0)) \quad (3.2)
\end{aligned}$$

Robins and Greenland [24] call the direct effect of the first decomposition “pure” and the direct effect of the second decomposition “indirect”. In the same way, they refer to the indirect effects of these two decomposition respectively as “total” and “pure”. These two ways of decomposing the total effect differ in the way in which they account for the interaction component. The total direct or indirect effect in fact incorporates also the interaction part. It is possible to choose a third decomposition in which the part due to interaction between the treatment and the exposure represents a part on its own. The decomposition is the following:

$$\begin{aligned}
Y(1) - Y(0) &= Y(1, m_1) - Y(0, m_0) \\
&= (Y(1, m_0) - Y(0, m_0)) + (Y(0, m_1) - Y(0, m_0)) + \\
&\quad (Y(1, 1) - Y(1, 0) - Y(0, 1) + Y(0, 0))(M(1) - M(0)) \quad (3.3)
\end{aligned}$$

The third term of the decomposition is called mediated interactive effect and it is equal to zero if the mediator has no effect on the outcome, that is:

$$M(1) - M(0) \neq 0 \quad (3.4)$$

or if there is no interaction between the treatment and the mediator variables, that is:

$$Y(1, 1) - Y(0, 0) \neq [(Y(1, 0) - Y(0, 0)) + (Y(0, 1) - Y(0, 0))] \quad (3.5)$$

Due to the fundamental problem of causal inference, it is impossible to estimate individual controlled and natural effects. However, under certain assumptions, we can estimate average effects. For example, we can identify an average controlled effect as:

$$E[Y(a, m) - Y(a^*, m)] = \sum_c \{E[Y | a, m, c] - E[Y | a^*, m, c]\} P(c) \quad (3.6)$$

In the same way, we can define average natural direct effect as:

$$E[Y(a, M_{a^*}) - Y(a^*, M_{a^*})] = \sum_c \sum_m \{E[Y | a, m, c] - E[Y | a^*, m, c]\} P(m | a^*, m, c) P(c) \quad (3.7)$$

and average natural indirect effects as:

$$E[Y(a, M_a) - Y(a, M_{a^*})] = \sum_c \sum_m E[Y | a, m, c] \{P(m | a, c) - P(m | a^*, c)\} P(c) \quad (3.8)$$

Assumptions

To estimate average controlled and natural effects, assumptions are needed, regardless of the measures chosen to quantify them.

The first assumption is that of no-unmeasured confounders for the relationship

between the exposure and the outcome and it is the same assumption we do in the general causal inference framework. Using mediation and counterfactual notation, we can formalize it in the following way:

$$Y(a, m) \perp A \mid C \quad (3.9)$$

and it holds for every c and for every m . This assumption is not sufficient to identify controlled effects: in this case, we should be sure that we have measured all the confounders for the mediator-outcome relationship. This assumption had been previously stressed by Jude and Kenny [14], but subsequent literature on mediation has omitted this issue. It is then necessary that the set of C variables contains all the covariates that confound the relationship between the mediator and the outcome. This assumption can be expressed as:

$$Y(a, m) \perp M \mid A, C \quad (3.10)$$

and it holds for every a and for every m . There are some contexts in which it is possible to identify mediator-outcome confounders that are affected by the exposure. Let L be the set of these variables. Thus, equation 3.10 modifies to:

$$Y(a, m) \perp M \mid A, C, L \quad (3.11)$$

These first two assumptions are sufficient for the estimation of controlled effects but for the identification of natural effects other two more assumptions are needed. First, we have to assume that all the exposure-mediator confounders have been measured that is:

$$M(a) \perp A \mid C \quad (3.12)$$

for all a levels. The conditions so far shown are no-unmeasured-confounding as-

sumptions. Thus, by collecting data on many variables, it is possible to hypothesize that they hold. To estimate natural direct and indirect effects, however, an additional assumption is required. In fact, it is necessary to assume that:

$$Y(a, m) \perp M(a^*) \mid C \quad (3.13)$$

for all a, a^* and for all m levels. This assumption is more difficult to understand but Pearl [27] showed that if condition holds and if it will generally be satisfied. Moreover, this fourth assumption can be omitted if assumptions 3.9 and 3.10 hold and if we assume that the effect of modifying the exposure level when fixing M to m value does not depend on m itself.

3.3 An illustration: odds ratio decomposition

An example of direct and indirect effect decomposition with a dichotomous outcome has been proposed by VanderWeele [28]. Suppose to have a dichotomous treatment variable A and let odds ratio (OR) be the measure that we choose to quantify the causal relationship with a treatment variable A . Then the total effect (TE), conditional on a vector of variables $C = c$, can be defined as:

$$OR^{TE} = \frac{P(Y(a) = 1 \mid c)/P(Y(a) = 1 \mid c)}{P(Y(a^*) = 1 \mid c)/P(Y(a^*) = 1 \mid c)} \quad (3.14)$$

This measure compares the odds of outcome $Y = 1$ when treatment is equal, respectively, to a and a^* . First, we can define a conditioned controlled direct effect (CDE) fixing variable M to a specific value m :

$$OR^{CDE}(m) = \frac{P(Y(a, m) = 1 \mid c)/P(Y(a, m) = 1 \mid c)}{P(Y(a^*, m) = 1 \mid c)/P(Y(a^*, m) = 1 \mid c)} \quad (3.15)$$

To decompose the total causal effect in direct and indirect components, we have to use concepts of natural direct and indirect effect. In the same way done by

Pearl [27] for risk difference measure, we can decompose the total causal OR into direct and indirect ORs. If there is not any significant interaction between treatment variable A and outcome Y then the controlled direct effect will be equal to the natural direct effect (NDE), otherwise they will be different for every fixed m .

The natural direct effect on the ratio scale can be defined as:

$$OR^{NDE} = \frac{P(Y(a, M_{a^*}) = 1 | c) / (1 - P(Y(a, M_{a^*}) = 1 | c))}{P(Y(a^*, M_{a^*}) = 1 | c) / (1 - P(Y(a^*, M_{a^*}) = 1 | c))} \quad (3.16)$$

In a similar way we can defined the natural indirect effect (NIE) on the odds ratio scale as:

$$OR^{NIE} = \frac{P(Y(a, M_a) = 1 | c) / (1 - P(Y(a, M_a) = 1 | c))}{P(Y(a, M_{a^*}) = 1 | c) / (1 - P(Y(a, M_{a^*}) = 1 | c))} \quad (3.17)$$

It can be shown that on the odds ratio scale, a multiplicative decomposition holds, that means that on the log odds ratio scale we have:

$$\log(OR^{TE}) = \log(OR^{NDE}) + \log(OR^{NIE}) \quad (3.18)$$

and this allows us to define a proportion of the mediated effect (PM) as:

$$PM = \frac{\log(OR^{NDE})}{\log(OR^{TE})}$$

If the assumptions stated in the previous paragraph hold, then we can estimate natural direct and indirect ORs. Suppose that the following model for the outcome holds:

$$\text{logit}(P(Y = 1 | a, m, c)) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta'_4 c \quad (3.19)$$

and the following model for the mediator holds:

$$E[M | a, c] = \beta_0 + \beta_1 a + \beta_2' c \quad (3.20)$$

where the error is supposed to be with mean equal to 0 and variance equal to σ^2 . Then natural direct and indirect log ORs can be estimated as:

$$\log(OR^{NIE}) \approx (\theta_2 \beta_1 + \theta_3 \beta_1 a)(a - a^*) \quad (3.21)$$

$$\log(OR^{NDE}) \approx \left\{ \theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 \sigma^2) \right\} (a - a^*) + 0.5 \theta_3^2 \sigma^2 (a^2 - a^{*2}) \quad (3.22)$$

where \approx holds for rare events and where σ^2 is the variance of the gaussian error term in model 3.20. The proof can be found in Vanderweele [28].

After point estimation, it is possible to estimate confidence interval for the controlled, direct and indirect effects [28]. Suppose to have estimated parameters of models 3.20 and 3.19 and suppose that the estimates have covariances matrices:

$$\Sigma_\beta = \begin{pmatrix} \sigma_{00}^\beta & \sigma_{01}^\beta & \sigma_{02}^\beta \\ \sigma_{10}^\beta & \sigma_{11}^\beta & \sigma_{12}^\beta \\ \sigma_{20}^\beta & \sigma_{21}^\beta & \sigma_{22}^\beta \end{pmatrix} \quad (3.23)$$

$$\Sigma_\theta = \begin{pmatrix} \sigma_{00}^\theta & \sigma_{01}^\theta & \sigma_{02}^\theta & \sigma_{03}^\theta & \sigma_{04}^\theta \\ \sigma_{10}^\theta & \sigma_{11}^\theta & \sigma_{12}^\theta & \sigma_{13}^\theta & \sigma_{14}^\theta \\ \sigma_{20}^\theta & \sigma_{21}^\theta & \sigma_{22}^\theta & \sigma_{23}^\theta & \sigma_{24}^\theta \\ \sigma_{30}^\theta & \sigma_{31}^\theta & \sigma_{32}^\theta & \sigma_{33}^\theta & \sigma_{34}^\theta \\ \sigma_{40}^\theta & \sigma_{41}^\theta & \sigma_{42}^\theta & \sigma_{43}^\theta & \sigma_{44}^\theta \end{pmatrix} \quad (3.24)$$

$$\Sigma_{\sigma^2} = (\sigma_{11}^{\sigma^2}) \quad (3.25)$$

then it is possible to use Delta Method and compute standard errors of the log of the controlled direct effect and of the natural indirect and direct effects as:

$$\sqrt{\Gamma\Sigma\Gamma'} | a - a^* | \quad (3.26)$$

where for the log of the controlled direct, natural indirect and natural direct effects, it holds, respectively:

$$\Gamma \equiv (0, 0, 0', 0, 1, 0, m, 0', 0)$$

$$\Gamma \equiv (0, \theta_2 + \theta_3 a, 0', 0, 0, \beta_1, \beta_1 a, 0', 0)$$

$$\Gamma \equiv (\theta_3, \theta_3 a^*, \theta_3 c, 0, 1, \theta_3 \sigma^2, \beta_0 + \beta_1 a^* + \beta_2' c + \theta_2 \sigma^2 + \theta_3 \sigma^2 (a + a^*), 0', \theta_3 \theta_2 + 0.5 \theta_3^2 (a + a^*))$$

and where Σ is defined as:

$$\Sigma = \begin{pmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_\theta & 0 \\ 0 & 0 & \Sigma_{\sigma^2} \end{pmatrix}$$

After computing standard errors for the log of the effect, it is straightforward to compute the confidence interval for the effect, by exponentiating the upper and lower extremes obtained for the log scale measures. It is worth noting that standard errors can be computed also through other methods, like bootstrapping. In the same way we can proceed for estimating confidence interval for the mediated proportion.

Chapter 4

Causal Mediation Analysis With Survival Data

4.1 Total, direct and indirect effect

In the causal mediation framework, decomposition with survival outcomes has been formalized. However, only few papers are available in the survival analysis literature, mainly from Lange [29] and VanderWeele [30]. In this Chapter we will present their approach.

Let A be a treatment, T a time-to-event outcome, M a mediator, and C a set of variable. We denote with T_a the counterfactual event time if A had taken value a and with T_{am} the counterfactual event time if A had taken value a and M had taken value m . As we previously did, we can define the following counterfactual $T_{aM_{a^*}}$, which denotes the event time when the exposure is set to a , but the mediator is set to the value it would have had if the exposure had been set to a^* . Again, we state composition assumption, that is $T_{aM_a} = T_a$. For a time-to-event variable V , we denote with $S_V(t) = P(V \geq t)$ the survival function at time t . Likewise, we define the the survival function conditional on covariates C as $S_V(t|c) = P(V \geq t|c)$. We use $\lambda_V(t)$ and $\lambda_V(t|c)$ for the hazard and conditional hazard at time t .

When dealing with survival data, different decomposition of a total effect into

direct and indirect effects are possible, in relation to the adopted scale. For example, it is possible to show that on the hazard scale, total effect can be decomposed in natural direct and indirect effects in the following way:

$$\begin{aligned} \lambda_{T_a}(t) - \lambda_{T_{a^*}}(t) &= [\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t)] + \\ &\quad [\lambda_{T_{a^*M_a}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)] \end{aligned} \quad (4.1)$$

where, the first expression is the indirect effect and the second the direct effect on the hazard scale. Moreover we could have chosen a three-way decomposition to highlight pure direct effect, pure indirect effect and the component due to the interaction between the treatment and the mediator. In this case, the decomposition would have been:

$$\begin{aligned} \lambda_{T_a}(t) - \lambda_{T_{a^*}}(t) &= [\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)] + \\ &\quad [\lambda_{T_{a^*M_a}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)] + \\ &\quad [\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t)] - \\ &\quad [\lambda_{T_{a^*M_a}}(t) - \lambda_{T_{a^*M_{a^*}}}(t)] \end{aligned} \quad (4.2)$$

In the same way, we could consider decomposition on the difference scale in terms of survival functions or in terms of mean survival times, as follows:

$$\begin{aligned} S_{T_a}(t) - S_{T_{a^*}}(t) &= [S_{T_{aM_a}}(t) - S_{T_{aM_{a^*}}}(t)] + \\ &\quad [S_{T_{aM_{a^*}}}(t) - S_{T_{a^*M_{a^*}}}(t)] \end{aligned} \quad (4.3)$$

$$\begin{aligned} E(T_a) - E(T_{a^*}) &= [E(T_{aM_a}) - E(T_{aM_{a^*}})] + \\ &\quad [E(T_{aM_{a^*}}) - E(T_{a^*M_{a^*}})] \end{aligned} \quad (4.4)$$

In certain contexts, decomposition on the log-hazard scale can be useful. In this case we have:

$$\begin{aligned} \log \{ \lambda_{T_{aM_a}}(t) \} - \log \{ \lambda_{T_{a^*M_{a^*}}}(t) \} = & [\log \{ \lambda_{T_{aM_a}}(t) \} - \log \{ \lambda_{T_{a^*M_{a^*}}}(t) \}] + \\ & [\log \{ \lambda_{T_{a^*M_{a^*}}}(t) \} - \log \{ \lambda_{T_{a^*M_{a^*}}}(t) \}] \end{aligned} \quad (4.5)$$

which can also be expressed as:

$$\begin{aligned} \lambda_{T_{aM_a}}(t) / \lambda_{T_{a^*M_{a^*}}}(t) = & \lambda_{T_{aM_a}}(t) / \lambda_{T_{a^*M_{a^*}}}(t) \times \\ & \lambda_{T_{a^*M_{a^*}}}(t) / \lambda_{T_{a^*M_{a^*}}}(t) \end{aligned} \quad (4.6)$$

The advantage of decompositions on difference scale is that it make possible to compute a mediation proportion by taking a ratio of the natural indirect effect to the total effect. Of course, measures of the proportion mediated may vary across scales. Moreover, the decomposition into direct and indirect effects may be analytically tractable on certain scales but not on others, depending on the survival model chosen.

Before proceeding and identifying direct and indirect effects we have to state the same assumptions of non-survival context. We then have to assume *i*) no unmeasured confounding of the exposure-outcome relation *ii*) no unmeasured confounding of the mediator-outcome relation *iii*) no unmeasured confounding of the exposure-mediator relation *iv*) no variable that is affected by *A* and that affects both *M* and *T*. This assumption then ensures that the exposure affects the outcome through two different pathways.

4.2 Aalen additive hazard model

The Aalen additive hazard model is a flexible semiparametric model for survival data. Contrary to the widely used Cox model, Aalen models the rate as a func-

tion of exposure A, mediator M and other baseline variables C in the following way:

$$\lambda(t) = \lambda_0(t) + \lambda_1(t)a + \lambda_2(t)c + \lambda_3(t)m \quad (4.7)$$

where λ_j are time-dependent coefficient functions. Moreover, we assume the mediator M to be normal and to be related to exposure A and covariates C through the following classic linear regression model:

$$M = \alpha_0 + \alpha_1 a + \lambda_2 c + e \quad (4.8)$$

where e is a gaussian error with mean equal to 0 and variance equal to σ^2 .

The indirect effect is the number of deaths that can be attributed to mediation through the mediator, whereas the direct effect is the number of deaths that can be attributed to a direct path (or to other mediators not included in the analysis). The total effect is the number of deaths caused by changing the exposure and it is equal to the sum of the direct and indirect effects.

If the assumptions (i)–(iv) hold, it can be shown that direct and indirect effects can be expressed in the following way:

$$\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t) = \lambda_1(t)(a - a^*) \quad (4.9)$$

$$\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t) = \lambda_3(t)\alpha_1(a - a^*) \quad (4.10)$$

When exposure and mediator are supposed not to be time-dependent then equations can be substituted by:

$$\lambda_{T_{aM_a}}(t) - \lambda_{T_{aM_{a^*}}}(t) = \lambda_1(a - a^*) \quad (4.11)$$

$$\lambda_{T_{aM_{a^*}}}(t) - \lambda_{T_{a^*M_{a^*}}}(t) = \lambda_3\alpha_1(a - a^*) \quad (4.12)$$

It possible to get estimates of direct and indirect effects and of the mediation proportion replacing λ_1 , λ_3 and α_1 in equations by their estimates $\hat{\lambda}_1$, $\hat{\lambda}_3$ and $\hat{\alpha}_1$. Under regular assumptions the random variables generated by these estimates are asymptotically normally distributed with the following covariance matrix:

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & 0 \\ \omega_{21} & \omega_{22} & 0 \\ 0 & 0 & \omega_{33} \end{pmatrix} \quad (4.13)$$

All elements of the covariance matrix are available from standard software. Direct effect is then asymptotically normal and also the indirect effect, being the product of two uncorrelated normal random variables. It follows that also total effect is normally distributed. Confidence interval for direct effect is immediately available, while for the indirect and total effect and for the mediation ratio confidence intervals can be obtained using delta method or through simulations.

The limitations of this method is that we have assumed a normal distributed mediator and we have made the assumption of no exposure-mediator interaction.

4.3 Cox proportional hazard model

Decomposition in natural indirect and direct effects is still possible when using well known Cox proportional hazard model: in this case, additive decomposition can not be done on the hazard scale but on the log-hazard one. When using Cox survival model, decomposition can account also for interaction between the exposure and the mediator variables. However, this decomposition is valid only if the studied outcome is rare, that is if $\lambda_T(t | 0, 0, 0)$ is small.

Thus, if we consider Cox survival model:

$$\lambda_T(t | a, m, c) = \lambda_T(t | 0, 0, 0)e^{\gamma_1 a + \gamma_2 m + \gamma_3 am + \gamma_4' c} \quad (4.14)$$

and if we suppose the following linear regression model for the mediator to hold:

$$M = \beta_0 + \beta_1 a + \beta_2' c + \epsilon \quad (4.15)$$

where the error ϵ is supposed to be follow a normal distribution with mean equal to 0 and variance equal to σ^2 , then we can decompose total effect into direct and indirect effects on the log-hazard scale as follows:

$$\log \{ \lambda_{T_{aM_a}}(t | c) \} - \log \{ \lambda_{T_{aM_{a^*}}}(t | c) \} = (\gamma_2 \beta_1 + \gamma_3 \beta_1 a)(a - a^*) \quad (4.16)$$

$$\begin{aligned} \log \{ \lambda_{T_{aM_{a^*}}}(t | c) \} - \log \{ \lambda_{T_{a^*M_{a^*}}}(t | c) \} &= \{ (\gamma_1 + \gamma_3(\beta_0 + \beta_1 a^* + \beta_2' c) \gamma_2 \sigma^2) \} \\ &\quad (a - a^*) + 0.5 \gamma_3^2 \sigma^2 (a^2 - a^{*2}) \end{aligned} \quad (4.17)$$

This means that the ratio between and the sum of and represents a measure of the proportion of the effect of the exposure mediated by the intermediate on the log-hazard scale. Again, we need simulations to get confidence intervals for the indirect and total effects and for the mediation proportion.

Decomposition without interactions

If there is not exposure-mediator interaction and the outcome is rare, it can be shown that estimates of direct and indirect effects are equal to the effects we would have estimated through difference or product method, whose results coincide in this particular context. This means that if we estimate the following model:

$$\lambda_T(t | a, m, c) = \lambda_T(t | 0, 0, 0) e^{\gamma_1 a + \gamma_2 m + \gamma_4' c} \quad (4.18)$$

and then we estimate the same model, omitting the mediator variable M :

$$\lambda_T(t | a, c) = \lambda_T(t | 0, 0) e^{\theta_1 a + \theta_4' c} \quad (4.19)$$

then the indirect effect can be computed as $\theta_1 - \gamma_1 = \gamma_2\beta_1$ and this holds for rare outcome. The proof of this can be found in Vanderweele [30].

4.4 Accelerated time failure model

Suppose to have the following accelerated time failure model, where the exposure, the mediator and an interaction between them are present:

$$\log(T) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 a m + \theta'_4 c + \nu \epsilon \quad (4.20)$$

and to have the following model for the mediator:

$$M = \beta_0 + \beta_1 a + \beta'_2 c + \epsilon \quad (4.21)$$

where ϵ is supposed to be normally distributed with mean equal to 0 and variance equal to σ^2 then the natural direct and indirect effect, conditional on $C = c$,

$$\log \{E(T_{aM_a})(t | c)\} - \log \{E(T_{aM_{a^*}})(t | c)\} = (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*) \quad (4.22)$$

$$\begin{aligned} \log \{E(T_{aM_{a^*}})(t | c)\} - \log \{E(T_{a^*M_{a^*}})(t | c)\} = & \{(\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta'_2 c))\theta_2\sigma^2\} \\ & (a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2}) \end{aligned} \quad (4.23)$$

Decomposition without interaction

Suppose the following model assuming no interaction between the exposure and the mediator can be used:

$$\log(T) = \theta_0 + \theta_1 a + \theta_2 m + \theta'_4 c + \nu \epsilon \quad (4.24)$$

and suppose also that the same model holds when there is no mediator variable, that is:

$$\log(T) = \theta_0 + \theta_1 a + \theta_2 M + \theta'_4 c + \kappa \epsilon \quad (4.25)$$

then product and difference methods gave same results. In fact, it is true that:

$$E[\log(T)|a, c] = \theta_0 + \theta_1 a + \theta'_4 c + \kappa E[\epsilon] \quad (4.26)$$

Moreover:

$$\begin{aligned} E[\log(T)|a, c] &= E[E[\log(T)|a, M, c]] \\ &= \theta_0 + \theta_1 a + \theta_2 E[M|a, c] + \theta'_4 c + \nu E[\epsilon] \\ &= \theta_0 + \theta_1 a + \theta_2 \{\beta_0 + \beta_1 a + \beta'_2 c\} + \theta'_4 c + \nu + E[\epsilon] \\ &= \{\theta_0 + \theta_2 \beta_0\} + \{\theta_1 + \theta_2 \beta_1\} a + \{\theta'_4 + \theta \beta'_2\} c + \nu + E[\epsilon] \quad (4.27) \end{aligned}$$

since this relation holds for every a , then we have $\theta_1 = \{\theta_1 + \theta_2 \beta_1\}$ and then $\theta_1 - \gamma_1 = \gamma_2 \beta_1$.

Chapter 5

Prostate Cancer

5.1 Epidemiology

Prostate cancer is one of the most common solid tumor in Europe and its incidence rate is 214 cases per 10.000 men every year. Prostate cancer affects more often elderly men then it has become a serious health concern in developed countries. Data show that 15% of male cancers are prostate cancer in developed countries compared with 4% of male cancers in developing countries. There are large regional differences in incidence rates of prostate cancer with a range from 68.8 in Malta to 182 in Belgium [31].

The factors that increase the risk of developing prostate cancer are not well known, although three well-known risk factors have been identified that is increasing age, ethnicity, and heredity.

If one first-line relative has the disease, the risk is doubled. If two or more first-line relatives are affected, the risk increases 5 to 11 fold [32]. About 9% of individuals with prostate cancer have true hereditary prostate cancer, defined as three or more relatives affected or at least two who have developed the disease before the age of 55 years.

Prostate cancer screening

There is currently no evidence for introducing widespread population-based screening programs for early prostate cancer detection. To evaluate the efficacy

of prostate cancer screening, two large randomized trials have been published: the Prostate, Lung, Colorectal, and Ovary (PLCO) trial in the United States and the European Randomized Study of Screening for Prostate Cancer (ERSPC) in Europe [33, 34].

The PLCO cancer screening trial assigned 76,693 men to receive either annual screening with prostate-specific antigen (PSA) and digital rectal examination (DRE) or standard care as the control. The follow-up was 7 years long and the incidence of prostate cancer per 10,000 person-years was 116 in the screening group and 95 in the control group. The incidence of death per 10,000 person-years was 2.0 in the screened group and 1.7 in the control group. The PLCO project team concluded that prostate cancer related mortality in screening detected individuals was not significantly different between the two study groups. The ERSPC trial included a total of 162,243 men between 55 and 69 years of age. The men were randomly assigned to a group offered a screening or to an unscreened control group. During a median follow-up of 9 years, the cumulative incidence of prostate cancer was 8.2% in the screened group and 4.8% in the control group. This resulted in an absolute risk difference of 0.71 deaths per 1,000 men. This means that 1,410 men would need to be screened and 48 additional cases of prostate cancer would need to be treated to prevent 1 death from prostate cancer. The ERSPC investigators concluded that PSA based screening reduced the rate of death from prostate cancer by 20% but was associated with a high risk of over-treatment.

Both trials have received considerable attention and comments. In the PLCO trial, the rate of compliance in the screening arm was 85% for PSA testing and 86% for DRE. However, the rate of contamination in the control arm was as high as 40% in the first year and increased to 52% in the sixth year for PSA testing and ranged from 41% to 46% for DRE. Furthermore, biopsy compliance was only 40.52% versus 86% in the ERSPC. Thus the PLCO trial will probably never be able to answer whether or not screening can influence prostate cancer mortality.

In a recent retrospective analysis of prostate cancer incidence, prostate cancer metastasis and cause of death were evaluated for a group of 11,970 men who were included in the intervention arm of the ERSCP trial and a control population of 133,287 unscreened men during an 8 years observation period [31]. The relative risk of prostate cancer metastasis in the screened population compared with the control population was 0.47 (p-value < 0.001). The RR of prostate cancer-specific mortality was also significantly lower in the screening arm. The absolute mortality reduction was 1.8 deaths per 1,000 men. Based on these data, the real benefit of the ESRPC trial will only be evident after 1015 years of follow-up, especially because the 41% reduction of metastasis in the screening arm will have an impact. Furthermore, time is needed to assess the economic burden and the side effects resulting from more intensive screening. Based on the results of these two large randomized trials, most if not all of the major urologic societies have concluded that at present widespread mass screening for prostate cancer is not appropriate. Rather, early detection should be offered to the well-informed man. Two key questions remain open and empirical: a) at what age should early detection start? b) what is the interval for PSA and DRE? The decision to undergo early PSA testing should be a shared decision between the patient and his physician based on information balancing its advantages and disadvantages. A baseline PSA determination at 40 years of age has been suggested on which the subsequent screening interval may then be based. A screening interval of 8 years might be enough in men with initial PSA levels 1 ng/ml. Further PSA testing is not necessary in men with age >75 years and a baseline PSA 3 ng/ml because of their very low risk of dying from prostate cancer [35].

Diagnosis of prostate cancer

The main diagnostic tools to diagnose prostate cancer include DRE, serum concentration of PSA, and transrectal ultrasound guided biopsies. In about 18 % of all patients, prostate cancer is detected by a suspect DRE alone, irrespective of the PSA level. A suspect DRE in patients with a PSA level of up to 2 ng/ml has

a positive predictive value of 530% [36]. A threshold level of PSA that indicates the highest risk of prostate cancer needs to be defined.

Method comparisons between the traditionally calibrated Hybritech PSA and free PSA (fPSA) assays and the new “standardised” World Health Organisation calibrated Access assays yielded results that are approximately 25% lower for PSA and fPSA. A PSA cut-off of 3 or 3.1 mg/l should be considered for WHO-calibrated assays to achieve the same sensitivity/ specificity profile as with a cut-off of 4 mg/l in traditionally calibrated assays.

TRUS or a transperineal laterally directed 18-gauge core biopsy has become the standard way to obtain material for histopathologic examination. The need for prostate biopsies should be determined on the basis of the PSA level and/or a suspicious DRE. The patients biologic age, potential comorbidities, and the therapeutic consequences should also be considered. The first elevated PSA level should not prompt an immediate biopsy, but it should be verified after a few weeks by the same assay under standardized conditions except for high PSA values >20 ng/ml once prostatitis has been excluded.

The Gleason score is recommended for grading prostate cancer. According to current international convention, the Gleason score of cancers detected in a prostate biopsy consists of the Gleason grade of the most extensive carcinoma component plus the highest grade, regardless of its extent (no 5% rule). In radical prostatectomy specimens, both the primary and the secondary Gleason grade should be reported. The presence of the tertiary grade and its approximate proportion of the cancer volume should also be reported.

Staging of prostate cancer

The TNM classification is used to stage prostate cancer. It describes the extent of the primary tumor (T stage), the absence or presence of spread to nearby lymph nodes (N stage) and the absence or presence of distant spread, or metastasis (M stage). The clinical stage is determined from information that is available without surgery. The pathologic stage is based on the surgical removal and histological examination of the entire prostate gland, the seminal vesicles and

surrounding structures and, if relevant, pelvic lymph nodes. The management of prostate cancer will depend on the TNM stage of the disease as well as both biochemical information, like PSA, and pathological information, like Gleason score, which have prognostic value. The optimum treatment for a man with prostate cancer requires an assessment of the risk of metastatic spread as well as the risk of local recurrence. For this, the results of imaging can be assessed in the light of information from clinical nomograms.

A nomogram is a statistically derived tool which is used to describe the likely course of a disease using known variables such as diagnostic findings, age and treatment options. Nomograms have been developed from outcome data on large groups of men with prostate cancer. Using predictive factors such as T-stage, Gleason score, PSA and histology results they can be used to estimate the risk of metastatic spread, lymph node involvement or recurrence following treatment. There is a wide variation in incidence rates between countries so that a nomogram developed in a screened population in the USA may not be wholly relevant to an unscreened population in an other country and therefore need to be used with caution.

Men newly diagnosed with prostate cancer can initially be stratified into those for whom radical treatment is a possibility and those for whom it is not appropriate. The decision about treatment intent will be based on the mans life expectancy, his values, and the anticipated clinical course of the prostate cancer.

High vs. low risk prostate cancer

Several factors have been shown to predict the risk of recurrence after treatment of prostate cancer as the Gleason score, the serum PSA level, and the T-stage. These predictive factors have been used to classify prostate cancer into risk groups. Different criteria are present in literature regarding how to do this classification. The assessment of risk at the diagnosis based on clinical data is fundamental since on that depends then the choice of the treatment. Since not all the diagnosed prostate cancers need to receive a treatment, it is crucial to understand which cancers need to be treated and which ones need only active

surveillance.

Active surveillance identifies a subgroup of men with prostate cancer expectantly with curative intent. Men are then carefully selected and subsequently actively observed in order to receive deferred curative treatment once the tumor seems to progress. Active surveillance is a strategy different from watchful waiting. Watchful waiting entails a strategy for all men who are managed expectantly, while active surveillance focuses on men for whom therapy is delayed until the tumor becomes progressive and curative treatment can be offered. This offers an attitude of active control over the cancer diagnosed for patients and their doctors. The stage migration that screening provides has resulted in an over-representation of low-risk cancers. Literature supports today the important role of active surveillance for men with low risk cancer. In this way, over-treatment and all the related side effects are avoided. Usually, patients with a Gleason score lower than 6, a clinical stage not higher than T2a and PSA lower than 10 ng/ml are classified as low-risk patients.

5.2 The role of physical activity

Physical activity plays a very important role in the prevention of a high number of diseases, improves quality of health and reduces the burden of chronic illness. Research and studies on cancer etiology conducted in the last 30 years have shown how physical activity and exercise can be important in the prevention of the development of tumors. The role of physical activity in reducing the risk of tumor has been studied for different type of cancer and the strongest evidence are relating breast and colorectal cancer.

To understand if physical activity is a factor that could have an impact on prostate cancer development, it is necessary first of all to understand natural history of prostate cancer. Natural history of this tumor is not well clear, however it is certain that prostate cancer has often a long phase in which the tumor is non symptomatic. Based on autopsy studies it has been estimated that in

the USA about 30% of men over 45 years of age has developed a latent prostate cancer. It has not been understood yet if it exists a type of prostate cancer that is entirely innocuous and it is also not clear if the factors that initiate the disease are the same that are responsible for the development of the disease from latent to clinical stage.

Even if many aspects of prostate cancer etiology are unclear, it has been proved that hormonal issues play an important role in determining its risk. Different levels of androgens can affect prostate gland functioning and its growth. Men who have very low levels of androgens, like castrated men, never develop prostate cancer and, in the same way, castration or estrogen therapy are important for the treatment of metastatic prostate cancer. Moreover, too high levels of androgens are associated with higher risk of developing prostate cancer. For example, black men in the USA have higher levels of testosterone than white men and have also higher risk of prostate cancer. Androgens could play a role in the initiation of the disease but hormone metabolism is involved also in the progression of the tumor. Since prostate cancer seems to have strong relations with hormonal issues, it is clear that one of the reasons for which physical activity could affect prostate cancer risk is through the hormonal changes. Indeed, physical activity induces changes for what regards hormone metabolism since it decreases levels of testosterone. For example, athletes have lower levels of testosterone and also persons who exercise experiment a temporary decrease in serum testosterone levels. However, how important can be the effect of these hormonal changes remains unclear.

Physical activity could have influence on the risk of prostate cancer also through its effect on body mass index (BMI). BMI has been found to be in turn an important and independent predictor of serum testosterone. BMI affects hormonal aspects since it has been shown that obesity leads to increment of testosterone levels due to a reduction of levels of sex-hormone binding globuline. Moreover, high levels of body fat stores results in increased androgens to estrogens conversion, that in turn could convert to dangerous carcinogens. Low fat in conjunction with a diet rich in fiber is effective in reducing circulating levels of

testosterone, thus physical activity, BMI control and a fiber diet can influence prostate cancer development.

So far, only biological mechanisms involving hormonal paths have been mentioned. However, physical activity could play a role in prostate cancer etiology also through other paths. Physical activity is indeed known to increase the number of natural killer cells and to improve immune function by inducing changes in the activity also of neutrophils and macrophages. Since relations between immune functions and cancer etiology have been studied, it is possible that physical activity could affect prostate cancer risk also through this path. Lastly, physical activity could affect prostate cancer etiology by affecting free radical levels. It has been shown indeed that free radical production could be increased by acute physical exercise, but chronic one should improve free radical defenses activating free radical scavenger enzymes.

The association between physical activity and prostate cancer was studied as well through large epidemiological studies. Results available in the medical literature do not converge to a unique answer. Two main reviews on this topic have been published: the first was published by Friedenreich and Thune [37] in 2001 and the second one was published by Young-McCaughan [38] in 2011. In their works all relevant publications were identified through Pubmed and results were summarized. Studies used different designs and among the considered studies there were prospective or retrospective cohort, case-control, nested case-control studies.

Each study assessed physical activity in a different way: through mailed questionnaires or direct interviews, subjects were asked to report information about their main daily occupation, the duration and intensity of recreational activities, type of performed sports, time spent in different daily activities. Some studies were able to quantify physical activity performed during life and evaluated then also the effect of exercise performed in the early stages of life.

Most of the studies quantified physical activity in Metabolic Equivalent of Task (MET): MET is a physiological measure expressing the energy cost of physical

activities and is defined as the ratio of metabolic rate (and therefore the rate of energy consumption) during a specific physical activity to a reference metabolic rate. MET are expressed in kcal/kg/hour. The MET concept was primarily designed for epidemiological studies in which it is required to quantify time spent for specific physical activities.

Friedenreich and Thune review conclusions were that “*the epidemiologic evidence is currently inconsistent and the magnitude of the risk reduction observed is small*”. Results obtained by Young through his review does not change consistently their previous results: risk reduction evidence remains small. However, Young results show that a trend suggesting a protective role of physical activity persists. Summarizing results from these two reviews, for a total of 40 studies, 55% found that physical activity reduces prostate cancer risk, 35% found that there is no significance association between the studied exposure and outcome and 10% showed an increase risk of prostate cancer due to physical activity. Moreover, in the 22 studies reporting a protective effect, 64% reported a small effect. In the same way, studies taht reported an increased effect, showed mainly small effects. It is also worth noting that evidence regarding physical activity decreasing prostate cancer risk is stronger when dealing with vigourous activity performed during adolescence or in the years immediately after.

Thus, the entity of the association between physical activity and prostate cancer has not been clarified yet but more than 30 years of research has shown that there could be a possible link between them.

5.3 The role of waist-to-hip ratio

Waist-to-hip ratio is a measure of central obesity, that has been suggested to be a potential risk factor for prostate cancer. Mechanisms through which increased central obesity might be linked to prostate cancer risk are still not well understood and several possible theories have been proposed. One hypothesis suggests that larger abdominal adiposity tends to lower the levels of free and total testos-

terone, as well as sex hormone binding globulin and that these alterations may, in turn, stimulate disease development. However, the influence of androgens alone does not fully explain the relationship, as evidenced in a study by Platz [39] who reported that an increase in total testosterone resulted in a decrease in the risk of high-grade prostate risk and in an increase in low-grade cancer risk. These findings suggest that the relations between the androgen pathway and prostate cancer is not direct and that disease pathogenesis is likely dependent also on tumor type and grade and other factors.

A second explanation for the underlying mechanisms related to prostate cancer development is based on the resulting effects of visceral adiposity on metabolic indicators as insulin resistance, regulation of leptin and lipid levels and release of free fatty acids [40]. These factors, alone or in conjunction, have been suggested to influence risk. Moreover, diet and lifestyle factors, hormonal changes, and genetic and environmental interactions play a role in the pathway between abdominal adiposity and prostate cancer. More research is needed to make clear the mechanisms contributing to the etiology of prostate carcinogenesis.

Results from studies analyzing the relation between waist-to-hip ratio and prostate cancer are not consistent. Several studies have indicated that waist-to-hip ratio is not associated with the development of prostate cancer, whereas others have reported a significant positive association [40]. Although some studies support the relationship between higher waist-to-hip ratio measurements and increased prostate cancer risk regardless of disease severity, the association has been shown to be stronger among those with advanced, high-grade disease [41, 42].

A large prospective study of more than 129,000 men from 8 European countries reported a significant association of waist-to-hip ratio with advanced prostate cancer but not low-grade cancer [41]. Similarly, the North Carolina Louisiana Prostate Cancer Project, which included 1,049 African-American men and 1,083 Caucasian-American males, found that men with a waist-to-hip ratio > 0.98 (compared with men with a waist-to-hip ratio < 0.90) had an increased risk of highly aggressive disease [42]. However, when the data were stratified by race,

the results showed that larger waist-to-hip ratio was not related to prostate cancer risk in men of African origin but was significantly associated with aggressive cancer in men of European descent.

Findings from the North Carolina Louisiana Prostate Cancer Project suggest potential differences in the relationship between body size and prostate cancer among different racial groups. The majority of studies were conducted in primarily European-derived populations, only one additional study assessed the relationship between waist-to-hip ratio and prostate cancer risk in a population of primarily African descent.

The hospital-based case control study conducted by Jackson and colleagues [43] found that Jamaican men with waist-to-hip ratio ≥ 0.95 had a double fold increased risk of high-grade disease compared with men in the reference range of waist-to-hip ratio and a significant association between waist-to-hip ratio and prostate cancer was reported among all cases, regardless of disease grade.

The Prostate Cancer in a Black Population study provides further support for an association between waist-to-hip ratio and prostate cancer risk, with more than a 2 fold increased risk for men in the fourth quartile of waist-to-hip ratio compared with those in the first quartile. This result was significant for all cases combined, as well as those in the high-grade subgroup alone.

It is worth noting that most of the investigations that found a positive association between waist-to-hip ratio and prostate cancer risk reported the finding only among men with advanced or aggressive disease. Interestingly, while most of these studies were conducted in European-derived populations, the 2 studies including Afro-Caribbean men found a significant relationship among all cases, regardless of disease severity.

It is unclear whether this discrepancy is the result of actual differences between Afro-Caribbean men and other groups or whether the finding is confounded by other factors [40].

Chapter 6

The Swedish Health Care System

6.1 General overview

Sweden has a population of around 9.000.000 subjects, almost exclusively Caucasian. Approximately 90% of the population resides on less than 30% of the area. More than 80% live in urban areas. Less than 5% of the working population are occupied within the agricultural sector. Wealth is unusually evenly distributed in Sweden: there are for instance no population strata that for economic reasons cannot get access to the best medical services available. The average life expectancy at birth is 78.8 for men and 83.6 for women: 1.53 million (17.2%) of the population are older than 64 years.

Since 1947, all Swedish residents have unique, 10-digit national registration number (NRN), which contains the date of birth and an additional 4 digits. The 9th digit can be used for gender identification. The 10th digit is a check sum which protects against incorrect data entries in computerized registers. The NRN is assigned to the individual immediately after birth, or after immigration. The NRNs are used extensively, both by official authorities and by health care, banks and businesses. Consequently, knowledge of the NRN is more or less a necessity for every Swedish resident and there are very few people who are

ignorant of their NRNs and who fail to give their correct NRNs upon request, e.g. in connection with health care visits.

Sweden is divided into 24 counties, the administrations of which bear responsibility for the maintenance of continuously updated, computerized registers, which in turn are compiled in central registers of the total population. In these registers, the NRNs are the unique identifiers. Since the administration of social insurance, health care, and pensions, among other things, is based on this registers, it is in the society's as well as in every person's own interest to keep the population registers updated. Thus, the registers receive immediate notifications upon birth, death, marriage, and relocation. Hence, the population registers are virtually 100% complete and entirely up-to-date.

With the exception of a small number of private practitioners who provide outpatient services, and a few very small private hospitals mainly devoted to minor surgery, the Swedish health care system is entirely public. It is organized by the county administrations under supervision of the National Board of Health and Welfare. Each county typically has 2-4 local hospitals and one county hospital, which serves as a first instance referral center and is capable of treating all but a few specialized cases. Basically, the catchment areas of the hospitals are mutually exclusive. The counties are organized in 6 health care regions, each containing one academic referral hospital. Technical facilities and management practices are uniform throughout the country, and there is practically no variation in health care quality between the counties. Every patient is obliged to use the hospital with the catchment area within which he/she resides (except for emergencies occurring outside the country). Thus, medical services are, in effect, population-based and referable to the county of residence.

All Swedish residents are covered by the mandatory social insurance, which reimburses the health care providers for all but a small part of the costs for both outpatient and inpatient care. Thus, Swedish health care is characterized

by excellent availability, high and uniform quality throughout the country, and low cost for patients. The fees are kept low enough to permit all residents equal access to public health care. There are no restrictions or barriers neither for socioeconomically unprivileged nor for ethnic minorities or women.

6.2 The Swedish National Registries

The Swedish Cancer Register

A nation-wide cancer register is in operation in Sweden since 1958. Both clinicians and pathologists/cytologists are required by law to notify the regional cancer register whenever a malignant condition is diagnosed, also autopsy. The notifications are checked extensively, including the NRNs, at the regional cancer registers. The national register is created annually at the National Board of Health and Welfare by merging of data from the 6 regional cancer registers. The cancer register is then linked to the death register to include dates and causes of death. In the register, all cancers are recorded with the coding system in current use along with a translation into ICD-7. Thus, until 1968 there was only ICD-7 coding, between 1968-1987 there was both ICD-7 and ICD-8 coding, and between 1987-1993 there was ICD-7 and ICD-9 coding. From 1993 onwards there is ICD-7 and ICD-0 coding. There is a special code for benign tumors, which are also reported to the register. There is a code for the basis for diagnosis, and a special code to indicate if a tumor was first detected at autopsy. Histological confirmation of tumors has varied over time and with tumor site, but is generally high and increasing with time: in a study of esophageal carcinoma, the percentage of cases verified by histology increased from 89% in 1960 to 98% in 1987, and in a study of cardiac cancer the percentage increased from 89% in 1970 to 97% in 1985. With respect to different histological cancer types, a special PAD coding and a SNOWMED classification have been in use since 1958.

The Cancer Register covers the entire Swedish population. Compliance with

the reporting requirements is excellent, and the register has been found to be more than 98% complete. With strictly equal access to health care for all population strata, including women and minorities, and almost complete reporting of incident cases to the cancer register, it can be safely concluded that the coverage is as complete for women and minorities with cancer as it is for all other population groups. Since the NRNs are carefully checked, the proportion with incorrect NRNs (including gender and birthdate) is negligible.

The Cancer Register provides cancer incidence data based on the experience of the entire general population, by calendar year, gender, five-year age-group, health care region, and site/type of cancer (on the three- or four-digit level of the ICD-7 classification). Further, corresponding incidence data can be generated for combinations of cancer site (according to ICD-7) and historical type (PAD code, see above). Incidence data can be obtained with or without inclusion of cases first detected at autopsy, and for first cancers only or for all cancers occurring in the population. For the former case, we have developed methods for estimating the prevalence pool of cancer in the Swedish population (which is not at risk and should be excluded from the denominator). Thus, we can provide exact estimates of the incidence of first cancers in Sweden. This information allows us to compute standardized incidence ratios in addition to relative risks based on internal comparisons within the cohort.

The population-based Inpatient Register

In 1964-1965, the National Board of Health and Welfare began collecting data on individual hospital discharges in the Inpatient Register. At discharge from hospitals, a specific form is completed for each patient, without exception. These forms are computerized locally, and the data are first stored in administrative registers held at the hospitals and the county administrations until delivery once a year to the National Board of Health and Welfare. Each record represents one in-hospital episode. In addition to the NRN and some administrative information, including admission and discharge dates, hospital and department

codes, and social insurance office code, it contains medical data such as up to 10 surgical codes (coded according to the Swedish Classification of Operations and Major Procedures), anesthesiological procedures, and 1-8 discharge diagnoses, coded according to the 10th revision of the International Classification of Diseases (ICD-10). The register has had nationwide coverage since 1987. Each year, the Inpatient Register includes approximately 1.7 million instances of hospital care.

As there are no restrictions with regard to availability of public health care services for any population group, and since all patients admitted to the participating hospitals are registered, there is no systematic underreporting of hospitalizations of any minority, and women of all age groups are included in the study population.

The proportion of patients with erroneous NRNs has varied with time, county, and diagnosis. This proportion was highest during the 1970s, and it has dropped substantially during more recent years. In a study performed by the National Board of Health and Welfare, the overall proportion of incorrect or missing NRNs was 7% in 1977, but it dropped to less than 2% in 1983. The proportion of records excluded due to obvious errors of the NRNs, NRNs that cannot be found in any other population register, or inconsistencies revealed during the linkage procedures which imply that the NRN is likely to be incorrect (date of birth after entry, death date before entry or before an in-hospital episode, different gender codes at successive hospitalizations) has ranged between 1 and 4% during the most recent years. In an analysis of a peptic ulcer cohort the percentage excluded was 2.4%, and in a study of hip replacement surgery, the proportion was 7%.

The proportion with invalid diagnostic codes or procedure codes (codes not listed in the code books) increased from 0.4% in 1964-1969 to 0.5-1.7% during the 1970s.

A few attempts have been made to quantify the underreporting to the Inpatient Register. The results are basically encouraging. In one study of hip fractures, underreporting was found to occur in less than 2% overall. In a study from 1994, the case records pertaining to a random sample of 900 in-hospital episodes were scrutinized, and the diagnoses made by an expert panel based on the case records were compared with the Inpatient Register data. In this stringent analysis, only main diagnoses were considered (thus, if the correct main diagnosis was entered as a co-diagnosis, this would be classified as incorrect coding in spite of the fact that such a patient would have been included in the correct cohort anyway). False negative cases of traumas, ischemic heart disease, and malignant tumors was found in 5, 7, and 3%, respectively. Underreporting of surgical procedures was noted in 8%, but in 3% these were only minor auxiliary procedures. False positive cases were found in 1, 2, and 2% among diagnoses of trauma, ischemic heart disease, and malignant tumors, respectively. In a recent small validation study of hemochromatosis, we found no false positive cases. The reasons for errors in the Inpatient Register are transfer errors in less than 1%, coding errors (correct diagnosis in the case record, but incorrect code given) in 6% and incorrect diagnoses in 8-11%.

The Death Register

The National Death Register is held by Statistics Sweden and is updated with about the same delay (2 years) as the Cancer Register. Besides date of death, the Death Register also holds information on underlying and contributory causes of death based on the International Classification of Disease codes.

For every deceased person a death certificate must be issued before burial is permitted. The death certificates are filled in by the physicians or surgeons in charge of the patient during the last hospitalization, or, for those few who die outside hospitals, by a family physician or a specialist in forensic medicine. Thus, there is no reason to believe that women or minorities are systematically excluded from the mechanism.

The national registration numbers are carefully checked by Statistics Sweden

before the data are entered into the register. The Death Register enables us to calculate expected numbers of deaths from specific causes by calendar year, gender, and 5-year age group.

The Register of the Total Swedish Population

This continuously updated register holds information on all individuals who are residents in Sweden. Linkage of the Inpatient Register to the Register of the total Swedish Population enables removal of all individuals not available for any follow-up from any particular cohort; they can be readily identified as those who are neither registered as dead or emigrated, nor as still alive and resident within Sweden. The Register of the Total Swedish Population is reliable and practically without erroneous NRNs.

The Register of Population Changes

Linkage to this register enables identification of all those individuals who have emigrated from Sweden to another country and in whom diagnosed cancer is thus unlikely to be reported to the Swedish Cancer Registry. When analyzing data, follow-up of these individuals is censored from the date of emigration.

The National Quality Registers

A system of National Quality Registries has been established in the Swedish health and medical services in the last decades. The National Quality Registries have been developed to fill the gap left by the lack of primary monitoring systems. The quality registries collect information on individual patients problems, interventions, and outcomes of interventions in a way that allows the data to be compiled for all patients and analyzed at the unit level. Since the registries are national, the entire country is in agreement on what indicates good care. This also makes it possible to compare different units. Currently there are 56 national quality registries in health care, 55 of which receive funds from the Executive Committee for National Quality Registries. In the areas where National Quality Registries have been established, the tools are available for

any unit that wants to participate to continuously monitor their effectiveness and the benefits that they create for patients.

The vision for the quality registries and the competence centers is to constitute an over-all knowledge system that is actively used on all levels for continuous learning, quality improvement and management of all healthcare services.

A catalog that describes the National Healthcare Quality Registries in Sweden was published 2007. The catalog also describes the organization of the Quality Registries and the Swedish Health Care system. A new catalog was published in Sweden 2010 but it is only in Swedish.

Eight competence centers for the National Quality Registries has been established. In a competence center, several registries share the costs for staff and systems that a single registry could not bear. Hence, a continued development of the registries can be assured although the system follows a decentralized model, i.e. each register is governed by a professional collaboration.

Competence centers aim to promote the development of new registries, create synergy effects by collaboration among registries (for example in technical operations, analytical work, and use of registry data to support clinical quality improvement), and helping to make registry data beneficial for different users.

6.3 National Prostate Cancer Register

Prostate cancer is one of the most common male cancer in Sweden, accounting for the 36% of new male tumors. Incidence of prostate cancer in Sweden has been increasing in the last decades, and the aged-standardized incidence rate doubled in the period 1990-2005. Mortality for prostate cancer has increased as well, even if not as quickly as incidence has done: in the period 1990-2004 the age-standardized rate passed from 67.2 to 71.5 every 100.000 persons years. Moreover, after 1980 an increasing number of localized prostate cancer were detected, due to the introduction of new detection methods. However, new diagnostic tools utilization varied largely across different part of Sweden. The higher number of detected localized prostate cancer also increased the devel-

opment of new curative treatments. Also new treatment utilization was not uniform around Sweden.

For these reasons, the need for urologists to have a database to collect data about new diagnosed prostate cancer became vital. This would have allowed also epidemiologists to evaluate if there was different incidence rate for different stages of the disease, to produce surveillance statistics and predict treatment outcomes. In the same time, urologists would have had an evidence based support for choosing the better treatment in relation to the characteristics of the patient.

In 1987 the first Prostate Cancer Register was set up: the register collected information about the characteristics of the primary tumor, the serum level of PSA and the treatment received by the patient. This register was set up in the south-east region of Sweden and in 1996 it was merged with three similar registers that after 1992 were set up in northern and southern regions of Sweden and in the regions of Uppsala/rebro. In 1997 and 1998 also the western and the Stockholm/Gotland region joined the register. The coverage of the register became than complete and National Prostate Cancer Register. National Prostate Cancer Register is one of the National Quality Register.

Today the register contains information regarding the following variables: date of diagnosis, diagnosing unit, cause of diagnosis, Gleason tumor grade, tumor stage according to the TNM classification, serum PSA level at the diagnosis, treatment received within 6 months after the diagnosis. Moreover, in the register are collected information also about the biopsy procedure and surgical treatment, as well as tumor extension in biopsies and surgical specimens. Regarding the TNM classification, only the major categories were used and no changes were made for them from the setting of the register.

New cases of prostate cancer are first reported by urologist or register nurses

to the respective regional register. Regional registers are then linked with regional cancer registers to identify missing cases and search for them. Every regional oncology center validates and checks for completeness data from regional prostate cancer registers. Once every year data are then reported to the National Prostate Cancer Register, based in the Uppsala/Orebro regional oncology center. Only from 2007 NPNs have been included in the National Prostate Cancer Register.

The National Prostate Cancer Register is administered by a team made up of two urologists from each of the Swedish healthcare regions, one oncologist, one pathologist and three members from three regional oncology centers.

The National Prostate Cancer Register is founded by the National Board of Health and Welfare and by the Association of the Swedish Counties. Every year the Register publishes online annual reports. Detailed information on the Register can be found in Adolfsson [44].

Chapter 7

Physical Activity and Prostate Cancer

7.1 Introduction

High levels of physical activity are associated with a reduced risk of several cancers including colon and breast [45]. Physical activity has been suggested to have a role also in the development of prostate cancer: possible biological mechanisms justifying this relation include hormonal changes induced by physical exercise. It has been shown that physical activity decreases serum testosterone levels and sex hormone binding globulin, that in turn can affect the risk of prostate cancer development [37]. However, studies assessing the relationship of physical activity with prostate cancer risk are inconclusive. Besides physical activity, abdominal adiposity also affects levels of testosterone and sex hormone binding globulin, increases levels of insulin and leptin that, in turn, modifies hormonal metabolism. Some studies have analyzed relationships between waist-to-hip ratio and prostate cancer risk. Results are inconsistent, but more evidence has been collected for aggressive prostate cancer [40].

Consequently, in a cohort of 13.000 Swedish men followed for 13 years, we investigated the association of physical activity and waist-to-hip ratio with prostate cancer and analyzed if waist-to-hip ratio could modulate the effect of physical activity on prostate cancer incidence/mortality. More in detail, we a) performed

analyses to quantify the causal relationship between different types of physical activity and prostate cancer risk b) analyzed the potential role of waist-to-hip ratio as a mediator in order to better understand the biological mechanisms from physical activity to prostate cancer.

Separated analyses were performed for total, low and high-risk prostate cancer.

7.2 Materials

The National March Cohort Data

In September 1997, the Swedish Cancer Society organized a four-day national fund-raising event, the National March, in almost 3.600 cities and villages around the country. In conjunction with this event, we established the Swedish National March Cohort. Participants in the march were expected to be particularly motivated to provide valid information. They were invited to fill out a 32 pages questionnaire.

Three pages were devoted to physical activity. Both previously validated and newly developed questions were used in order to estimate energy expenditure and to distinguish between constant low-level activity and short-term peak activity. We used a new rating scale for self-reports of time spent on different intensity levels of physical activity (and inactivity) during a typical day, allowing estimation of total energy expenditure on an interval scale level.

Seven pages contained questions about diet, using a 106-item validated semi-quantitative food frequency questionnaire that allows estimation of total energy intake and calorie adjustment. Supplementary questions were asked about intake of fried food, detailed pattern of alcohol intake, dietary supplements and use of health food preparations. Two pages had questions about anthropometric measures, including height, weight (birth weight, current weight, and weight fluctuations), waist and hip measures (allowing calculations of BMI and waist-to-hip ratio).

Further, three pages were allocated to questions about various background and possibly confounding factors such as country of birth, environment during child-

hood and adolescence, birthplace of parents, own education, type of employment.

Two pages were about smoking (including passive smoking) and snuff dipping habits. One page was dedicated to vaccination history, two pages to medical history (including questions about allergies, asthma, acne, type 1 and type 2 diabetes, hypertension, lipid abnormalities, myocardial infarction, angina pectoris, claudicatio intermittens, stroke, rheumatoid arthritis, cancer, MS, inflammatory bowel disease, fractures), and two pages to pharmacological history. Two pages were about sun and UV exposure and type of complexion.

Five pages were spent on questions concerning the psychosocial history, including validated sets of questions about demands and autonomy at work, life events, self-perceived health, social support, as well as duration and quality of sleep. Sleep disturbances were assessed using a modified version of the Karolinska Sleep Questionnaire (KSQ), a widely used tool to assess quality and restorative function of sleep. One page was about the use of mobile telephones. Two pages were devoted to questions specific for women (age at menarche, parity, infertility, menstruation, menopause, use of oral contraceptives and hormone replacement therapy).

Due to the open nature of the National March, the total number of individuals who in reality were given a questionnaire could not be assessed. In total, 43.880 men and women completed the questionnaire. Data cleaning led to an unusually low number of exclusions; four questionnaires were found to be returned totally incomplete and 13 had incorrect national NRNs. Moreover, upon the initial linkages to Swedish demographic registers, 4 and 11 individuals were shown to have dates of emigration and death, respectively, prior to date of enrollment, leaving 43.848 analyzable subjects in the National March Cohort. More information on the cohort and on the used instruments can be found in Lagerros [46].

Exposures, covariates and outcomes

Data collected during the National March Cohort were linked with the following national registers through the PNR: Statistics Sweden Register of the Total Population and Population Changes, Patient Register, Cancer Register, and Causes of Death Register. Finally, linkage with Prostate Cancer Quality Register was performed.

Baseline measurements

Exposure and covariate data were based on self-reported information collected in the questionnaire given at cohort enrollment. We collected information on total, occupational and leisure physical activity, waist and hip circumference, alcohol consumption, smoking status, educational level and diagnosis of diabetes. Waist-to-hip ratio was calculated as the ratio of waist circumference to hip circumference.

An index of frequency of physician contacts was created assessing if the subject during his life had contacted a doctor for one of the following reasons: allergic skin problems, allergic nasal catarrh, acne, contact eczema, psoriasis, asthma, heart attack, high blood pressure, angina pectoris, lipid disturbance, stroke, rheumatoid arthritis, tuberculosis, wrist fracture, multiple sclerosis, Crohns disease, ulcerative colitis. Tertiles were created on the basis of the number of diseases reported.

We created this index because a subjects who often contacts the healthcare system, has higher probability of being screened for other diseases.

Physical activity during a typical day was estimated using a validated Energy Expenditure Questionnaire (EEQ). EEQ comprise nine physical activity steps grading physical activity according to intensity levels. Each step illustrates common activities and is assigned a value expressed as a multiple of metabolic energy turnover (MET). MET, as defined earlier, is computed as the ratio of work metabolic rate to a standard resting metabolic rate, it is expressed in $kcal \cdot (kg \cdot hour)^{-1}$, with 1 MET equivalent to sitting quietly. The MET values

assigned to the nine different steps ranged between 0.9 and 8.0 METs. Participants were instructed to report the time spent on each intensity level during a typical day and night. Hence, the total physical activity time ought to add up to 24 hours, allowing for an estimate of MET-hours per day (MET_h/day).

Occupational physical activity was assessed through the question “How physically demanding has your daily work/occupation been during the past 12 months?”. Possible answers were “light, mostly sitting”, “light, moving around”, “rather physically demanding” and “very physically demanding”. An index of low/high muscular or locomotive activity during daily working hours was also created.

To assess leisure physical activity, participants were asked to report the time spent on different exercise/outdoor life activities in both summer and winter times during the previous 12 months. Participants were also asked how many hours they devoted every week in activities like walking and/or biking to work, by weekly cleaning, gardening or the alike. Each activity was assigned a MET value. Hence, leisure physical activity time was obtained as an average of MET-hours per day spent in leisure activities involving physical activity. Separated indexes of low/high light (taking a walk), moderate (speedy walking, jogging, swimming) and vigorous (hard training or competition) leisure physical activity were also created.

A continuous index of physical activity across the lifespan was computed assessing the number of times the subject had been exercising weekly during different ages (10-19, 20-29, 30-49 and more than 50 years old).

Follow-up

The cohort was followed-up for 13 years and 3 months, starting on October 1, 1997. When analyzing incidence data, follow-up ended at time of prostate cancer diagnosis, death due to prostate cancer, emigration or December 31, 2010, whichever occurred first. Incidence data were obtained through linkage with the Swedish National Cancer Register, estimated to be almost 100% complete.

Low-risk tumors were defined as having a T lower than 3, Gleason sum score lower or equal to 7 and serum PSA lower or equal to 20 ng/ml. Subjects who had T higher or equal to 3, or Gleason sum score higher than 7, or serum PSA higher than 20 ng/ml, or lymph nodes involvement, or presence of distant metastasis were classified as high-risk cancers.

When analyzing mortality data, follow-up ended at time of death for prostate cancer, death occurring for other causes, emigration or December 31, 2010, whichever occurred first. Mortality data were obtained by querying the more than 97% complete Swedish Death Registry.

7.3 Methods

Primary analyses

The distribution of total physical activity was categorized into low, medium and high levels by dividing METh/day into tertiles, with cutoffs at 34.1 and 45.5 METh.

Occupational physical activity was categorized into low, medium and high according to the fact that the subject had answered, respectively, “light, mostly sitting / light, moving around” or “rather physically demanding” or “very physically demanding” to the question regarding physical strain at work.

The distribution of leisure physical activity was categorized into low, medium and high levels by dividing METh/day concerning leisure activities into tertiles, with cutoffs at 1.5 and 3.2 METh.

Pearson’s chi-squared tests and one-way ANOVA were used to compare the differences between frequencies and means, respectively, among different levels of waist-to-hip ratio and different levels of total physical activity.

Cox regression models were fitted to estimate hazard ratios and 95% confidence intervals of prostate cancer incidence/mortality at different levels of waist-to-hip ratio, total, occupational and leisure physical activity. Low waist-to-hip ratio and low physical activity (total, occupational and leisure) were the referent categories. Interactions between total, occupational and leisure physical activity

and waist-to-hip ratio were assessed by including the cross-product interaction term with the main effect terms in the Cox regression model.

Statistical significance of interaction was obtained by comparing nested models with and without the introduction of the interaction term with the likelihood ratio test.

Ties were handled using the Breslow method. The proportional-hazards assumption in the Cox regression model was tested by using Schoenfeld residuals. In a first analysis we estimated age-adjusted hazard ratios. In a secondary analysis, the following potential confounders were adjusted for: alcohol drinking (non-drinkers, low-drinkers, high-drinkers), smoking status (never, former, current), level of education (<12 years, >12 years), diabetes (yes/no), frequency of physician contacts (low, medium, high).

Sensitivity analyses were performed modeling waist-to-hip ratio and physical activity (total and leisure) as continuous variables through restricted cubic splines. Sensitivity analyses were also performed excluding the first 2 and 4 years of follow-up. In this way we controlled for reverse causality, being physical inactivity potentially related to sub-clinical disease.

Mediation analysis

To understand if waist-to-hip ratio could be considered a mediator in the relation between physical activity and prostate cancer incidence/mortality we performed a mediation analysis following the theory developed by VanderWeele and presented in Chapter 4.

We assumed that waist-to-hip ratio measure follows temporally physical activity: this assumption was reasonable since the subject was asked about his physical activity in the previous 12 months, while waist-to-hip ratio was measured at the recruitment of the cohort.

Thus, we decomposed the total effect of physical activity on prostate cancer in a part that is due to the pathway through waist-to-hip ratio decrease and a part that it is not due to this pathway. We verified that the assumptions for estimating natural direct and indirect effects were satisfied and we assumed that

there was no unmeasured confounding.

Among the possible survival model, we chose proportional Cox model that allows a decomposition on the log-hazard scale. In this analysis waist-to-hip ratio variable was used not as a categorical one but as a continuous one. The theory developed by VanderWeele allowed to take into account potential interactions between waist-to-hip ratio and physical activity.

We reported two sided P-values. All analyses were performed using Stata release 12 (StataCorp, College Station, TX, USA). For mediation analysis we used in Stata a macro to compute punctual and interval estimation of total, direct and indirect effects.

7.4 Results

Descriptive analyses

After applying exclusion criteria, our initial cohort comprised 13.109 men with mean age equal to 55 years. Characteristics of the study population can be found in Table 7.1. We report levels of physical activity stratified by level of cohort characteristics: we observe that men in different total physical activity categories were different in terms of educational level, alcohol consumption, age, smoking status, frequency of physician contacts and physical activity. Men who were healthy, young and less educated seemed to be more physically active on average. In the same way, in Table 7.2 we report levels of waist-to-hip ratio stratified by other baseline variables: the distribution was not statistically equal among different classes and we observed that young and healthy men who did not smoke, drink alcohol and with higher education level had lower levels of waist-to-hip ratio.

Waist-to-hip ratio and prostate cancer

Results are displayed in Table 7.3. The interaction between total physical activity and waist-to-hip ratio was not statistically significant. Furthermore, there

Table 7.1: Descriptive statistics stratified by Total Physical Activity

	Total physical activity*			p-value
	low	medium	high	
Age at baseline (mean, years)	56.23	54.75	54.35	<0.001
Waist-to-hip ratio (mean)	0.94	0.94	0.93	<0.001
Education level (>12 years, %)	30.97	29.95	15.30	<0.001
Physician contact (%)				<0.001
low	44.60	49.07	51.17	
medium	28.84	29.02	27.39	
high	26.57	21.91	21.44	
Smoke (%)				<0.001
non smoker	57.81	60.87	62.67	
former	35.03	33.34	31.10	
current	7.16	5.79	6.23	
Diabetes (yes, %)	4.51	3.26	3.53	0.007
Alcohol (%)				<0.001
low	9.44	9.20	10.29	
medium	34.04	34.89	38.25	
high	56.52	55.91	51.46	

*Total pa. low (<34.2 METh/day) medium (>34.2 < 45.9 METh/day)
high (>45.9 METh/day)

Table 7.2: Descriptive statistics stratified by Waist-to-hip ratio

	Waist-to-Hip Ratio*			p-value
	low	medium	high	
Age at baseline (mean, years)	54.26	58.93	58.54	<0.001
Total physical activity (mean, MET _h /day)	44.50	43.37	41.90	<0.001
Education level (>12 years, %)	30.00	23.65	22.28	<0.001
Physician contact (%)				<0.001
low	52.23	46.83	43.31	
medium	28.27	29.48	28.38	
high	19.50	23.69	28.31	
Smoke (%)				<0.001
no	67.02	58.87	53.49	
former	28.00	34.98	40.71	
current	4.99	6.15	5.80	
Diabetes (yes, %)	2.61	3.76	5.23	<0.001
Alcohol (%)				<0.001
low	10.65	10.94	9.64	
medium	38.56	33.83	32.99	
high	50.80	55.23	57.37	

*Waist-to-hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

was no evidence of deviation from proportionality assumption. We observed a positive relationship between waist-to-hip ratio and prostate cancer: subjects with waist-to-hip ratio higher than 0.85 seemed to be at higher risk of developing prostate cancer. This result was seen both in the age and in the multivariable adjusted models; the estimated hazard ratio was statistically significant only for the category of men with waist-to-hip ratio lower than 0.90 in the multivariable adjusted model (HR 1.25; CI 1.02-1.54).

In a second analysis, we run the same adjusted model used before, but instead of adjusting for total physical activity, we considered the physical activity components due to daily work and to leisure exercise. Results are displayed in Table 7.4. Again, we observed the same trend and we obtained the significance of the results only for the second waist-to-hip ratio tertile (>0.85 ; <0.90) in the multivariable adjusted model (HR 1.41; CI 1.11-1.78).

Data from National Prostate Cancer Register allowed us to perform subgroups analyses for different type of tumors at diagnosis. We observed the same results for both low and high-risk prostate cancer. Statistical significance was achieved only for low-risk prostate cancer when considering second waist-to-hip ratio tertile (HR 1.49; CI 1.06-2.09). See Table 7.4.

Additionally, sensitivity analyses were performed and we modeled waist-to-hip ratio as continuous variables through restricted cubic splines. There were no significant trends.

Results presented in Table 7.3 and Table 7.4 show that subjects in the second waist-to-hip ratio tertile (>0.85 ; <0.90) had higher risk to die for prostate cancer and this higher risk is with both age and multivariable adjusted models than men in the lowest class (<0.85). This result, however, was not statistically significant. Moreover, we did not observe any trend and men in the last tertile did not have any increased risk compared with men with waist-to-hip ratio lower than 0.85. No statistically significant evidence of interaction between waist-to-hip ratio and total, occupational and leisure physical activity was found.

We repeated the analyses using restricted cubic splines for modeling waist-to-hip

ratio. Results did not change.

Physical activity and prostate cancer

We found that there was no association between total physical activity and total prostate cancer incidence. The absence of association was observed both when we adjust only for age and when we adjust for all the others confounders (age, education level, frequency of physician contacts, alcohol consumption, smoking status, diabetes).

When analyzing occupational physical activity and total cancer incidence, a different trend was found. We observed that higher levels of occupational physical activity were associated with a decrease in the risk of total prostate cancer incidence and this trend remained valid after adjustment for all the confounders considered in the analysis. Results were significantly significant when considering the highest level of occupational physical activity and only in the multivariable adjusted model (HR 0.70; C.I. 0.50-1.00).

No similar trend was observed when analyzing leisure physical activity. Interactions between physical activity (total, occupational and leisure physical activity) and waist-to-hip ratio were tested using likelihood ratio test, but they were not statistically significant.

Results did not change if we use physical activity (total and leisure) as continuous variables, modeled through restricted cubic splines. When analyzing low and high-risk prostate cancer separately, we observed no effect of leisure and total physical activity, but occupational physical activity seemed to be protective for low-risk cancer in the multivariable adjusted model (HR 0.60; C.I. 0.36-0.98).

We performed further analyses both for occupational and leisure physical activity. Regarding occupational physical activity, we analyzed separately the locomotion and muscular component and we observed strong evidence only for the former. The association was significant for both total (HR 0.62; C.I.0.42-0.89) and low-risk (HR 0.51; 0.29-0.89) prostate cancer.

We repeated the previous analyses for both occupational and muscular activity

after stratifying for level of education in order to ensure a better control for social class level. The same trends were observable for both high (>12 years) and low (<12 years) education, however significance held only when analyzing separately the effect of the muscular component in the low educated men. These results were valid for both total (HR 0.65; C.I. 0.44-0.96) and low-risk prostate cancer (HR 0.48; C.I. 0.27-0.92).

Regarding leisure physical activity, we analyzed separately the association with light, moderate and vigorous leisure physical activity and cancer incidence. We observed a non-significant reduction in prostate cancer incidence associated with high levels of light physical activity but a significant increase of risk associated with high levels of moderate physical activity for both total (HR 1.23; C.I. 1.02-1.49) and low-risk (HR 1.34; C.I. 1.03-1.74) cancer incidence.

Finally, we run a further analysis on the subgroup population for which age at the beginning of the study was higher than 30 years, to understand if leisure physical activity performed before the age of 30 could affect prostate cancer incidence. No associations were found (data not shown).

To avoid the risk of undiagnosed prostate cancer affecting physical activity, we repeated analyses excluding the first 2 and 4 years of follow-up. The most important difference that emerged in both these scenarios was the loss of the significance of the association between occupational activity and total and low-risk prostate cancer incidence.

We found that higher levels of total and occupational physical activity were associated with a decrease for prostate cancer mortality risk. Results are shown in Table 7.3 and Table 7.4. Results were not statistically significant and do not change when excluding the first 2 and 4 years of follow-up to avoid reverse causality bias. No associations are found with leisure physical activity performed during youth.

Table 7.3: Associations between total physical activity and waist-to-hip ratio and prostate cancer incidence/mortality

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality		
	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]
Tpa*	1,000			1,000			1,000			1,000		
low	1,061	0,564	0,868	1,207	0,196	0,907	1,605	0,643	0,691	0,913	0,735	0,538
medium	1,075	0,489	0,876	1,182	0,266	0,880	1,588	0,833	0,716	0,741	0,309	0,416
high												1,320
WHR**	1,000			1,000			1,000			1,000		
low	1,251	0,035	1,016	1,198	0,234	0,890	1,613	0,254	0,880	1,206	0,510	0,690
medium	1,219	0,068	0,986	1,292	0,092	0,959	1,740	0,403	0,834	0,969	0,917	0,534
high												1,756
Education	1,000			1,000			1,000			1,000		
<12 yrs	1,136	0,201	0,934	1,208	0,168	0,923	1,580	0,706	0,783	1,444	0,177	0,848
>12 yrs												2,461
Ph.cont.	1,000			1,000			1,000			1,000		
low	1,219	0,042	1,007	1,149	0,315	0,876	1,508	0,041	1,012	2,600	0,001	1,488
medium	1,042	0,702	0,842	1,070	0,658	0,793	1,443	0,991	0,721	1,540	0,181	0,818
high												2,901
Smoke	1,000			1,000			1,000			1,000		
no	1,042	0,649	0,873	1,041	0,749	0,812	1,335	0,697	0,807	0,960	0,874	0,579
former	1,327	0,127	0,923	1,400	0,171	0,865	2,268	0,355	0,738	2,091	0,126	0,812
current												5,384
Diabetes	1,000			1,000			1,000			1,000		
no	0,892	0,603	0,581	0,778	0,448	0,408	1,486	0,907	0,517	1,847	0,137	0,822
yes												4,151
Alcohol	1,000			1,000			1,000			1,000		
non	0,991	0,948	0,761	1,143	0,527	0,755	1,731	0,391	0,592	1,016	0,963	0,505
low	0,883	0,344	0,682	1,143	0,542	0,759	1,690	0,040	0,477	0,962	0,911	0,486
high												1,905

Abbreviations: Tpa= Total physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals.

Ph.cont.= Physician contact.

^aH.R. are age-adjusted *tpa. low (<34.2 METh/day) medium (>34.2 < 45.9 METh/day) high (>45.9 METh/day)

**waist to hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

Table 7.4: Associations between occupational and leisure physical activity and prostate cancer incidence/mortality

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality		
	H.R. ^a	P.V.	95% C.I.]	H.R. ^a	P.V.	95% C.I.]	H.R. ^a	P.V.	95% C.I.]	H.R. ^a	P.V.	95% C.I.]
Opa	1,000			1,000			1,000			1,000		
low	0,892	0,403	0,682	0,893	0,534	0,625	0,844	0,432	0,552	0,667	0,321	0,300
medium	0,703	0,048	0,495	0,599	0,043	0,364	0,765	0,319	0,451	0,388	0,110	0,121
high			0,997			0,984			1,296			1,241
Lpa*	1,000			1,000			1,000			1,000		
low	1,109	0,365	0,887	1,066	0,692	0,777	1,025	0,884	0,735	1,509	0,246	0,753
medium	1,006	0,961	0,787	1,074	0,678	0,766	0,877	0,487	0,606	1,474	0,321	0,684
high			1,285			1,507			1,269			3,176
WHR**	1,000			1,000			1,000			1,000		
low	1,405	0,005	1,111	1,488	0,022	1,058	1,132	0,475	0,805	1,326	0,411	0,677
medium	1,196	0,156	0,934	1,413	0,053	0,996	0,961	0,827	0,669	0,929	0,844	0,444
high			1,531			2,004			1,380			1,942
Education	1,184	0,123	0,955	1,220	0,186	0,909	1,107	0,552	0,792	1,529	0,179	0,823
<12 yrs			1,467			1,639			1,547			2,840
>12 yrs												
Ph.cont.	1,000			1,000			1,000			1,000		
low	1,237	0,052	0,998	1,186	0,270	0,876	1,359	0,060	0,987	2,412	0,008	1,253
medium	1,088	0,492	0,856	1,106	0,557	0,790	1,020	0,917	0,706	1,227	0,601	0,569
high			1,383			1,548			1,473			2,646
Smoke	1,000			1,000			1,000			1,000		
no	1,080	0,446	0,886	1,118	0,430	0,848	1,051	0,747	0,777	0,981	0,951	1,795
former	1,405	0,094	0,944	1,482	0,146	0,872	1,353	0,347	0,721	2,373	0,119	0,801
current			2,090			2,518			2,540			7,033
Diabetes	1,000			1,000			1,000			1,000		
no	0,962	0,872	0,597	1,006	0,986	0,520	0,925	0,834	0,445	2,128	0,132	0,797
yes			1,548			1,945			1,920			5,683
Alcohol	1,000			1,000			1,000			1,000		
low	0,927	0,623	0,686	1,123	0,630	0,700	0,772	0,222	0,510	1,124	0,799	0,457
medium	0,895	0,456	0,669	1,147	0,554	0,728	0,713	0,100	0,476	1,288	0,565	0,544
high			1,198			1,809			1,067			3,048

Abbreviations: Opa= Occupational physical activity; Lpa= Leisure physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals.
 Ph.cont.= Physician contact.
^aH.R. are age-adjusted *lpa. low (<1.5 METh/day) medium (>1.5 <3.3 METh/day) high (>3.3 METh/day)
 **waist to hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

Table 7.5: Associations between occupational (muscular activity) and leisure physical activity and prostate cancer incidence/mortality

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality		
	H.R. ^a	P.V.	C.I.	H.R. ^a	P.V.	C.I.	H.R. ^a	P.V.	C.I.	H.R. ^a	P.V.	C.I.
Oppa	1,000			1,000			1,000			1,000		
low	0,615	0,011	0,424	0,509	0,019	0,290	0,894			0,900	0,841	2,516
high												
Lpa*	1,000			1,000			1,000			1,000		
low	1,096	0,419	0,877	1,054	0,744	0,769	1,445			1,456	0,285	2,900
medium	0,985	0,905	0,772	1,044	0,802	0,745	1,462			1,344	0,445	2,867
high												
WHR**	1,000			1,000			1,000			1,000		
low	1,405	0,005	1,111	1,777	0,022	1,058	2,094			1,346	0,387	2,637
medium	1,196	0,155	0,935	1,532	0,051	0,999	2,011			0,941	0,872	1,969
high												
Education	1,000			1,000			1,000			1,000		
<12 yrs	1,199	0,093	0,970	1,483	0,144	0,928	1,663			1,626	0,120	3,001
>12 yrs												
Ph.cont.	1,000			1,000			1,000			1,000		
low	1,236	0,054	0,996	1,532	0,268	0,876	1,608			2,433	0,008	4,683
medium	1,084	0,510	0,853	1,378	0,577	0,786	1,540			1,255	0,561	2,705
high												
Smoke	1,000			1,000			1,000			1,000		
no	1,086	0,417	0,890	1,324	0,399	0,854	1,486			1,004	0,991	1,835
former	1,414	0,087	0,950	2,103	0,137	0,880	2,541			2,516	0,095	7,428
current												
Diabetes	1,000			1,000			1,000			1,000		
no	0,963	0,877	0,598	1,550	0,977	0,523	1,951			2,111	0,136	5,638
yes												
Alcohol	1,000			1,000			1,000			1,000		
low	0,927	0,624	0,686	1,254	0,618	0,703	1,809			1,115	0,812	2,745
medium	0,900	0,477	0,673	1,204	0,526	0,735	1,826			1,312	0,537	3,103
high												

Abbreviations: Oppa= Occupational physical activity; Lpa= Leisure physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals.

Ph.cont.= Physician contact.

^aH.R. are age-adjusted *lpa. low (<1.5 METH/day) medium (>1.5 <3.3 METH/day) high (>3.3 METH/day)

**waist to hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

Table 7.6: Associations between different type of leisure physical activity and prostate cancer incidence/mortality

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality		
	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]
Light pa low high	1,000			1,000			1,000			1,000		
	0,906	0,086	0,752	1,092	1,039	1,301	0,991	0,946	0,754	1,101	0,704	1,808
Mod. pa low high	1,000			1,000			1,000			1,000		
	1,232	0,032	1,019	1,489	1,737	1,458	1,085	0,591	0,806	1,064	0,838	1,929
Vig. pa low high	1,000			1,000			1,000			1,000		
	0,981	0,870	0,776	1,239	1,490	1,300	0,899	0,572	0,622	0,862	0,715	1,911

Abbreviations: pa=physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals; Mod=moderate; Vig=vigorous
^aH.R. are age-adjusted are adjusted for the following variables: baseline age, occupational pa, waist-to-hip ratio, education level, physician contact, smoke, diabetes, alcohol

Table 7.7: Associations between occupational and leisure physical activity and prostate cancer incidence/mortality, low education level

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality		
	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]
Op low	1,000			1,000			1,000			1,000		
medium	0,959	0,815	0,678	0,959	0,861	0,602	0,856	0,564	0,506	0,375	0,029	0,155
high	0,801	0,289	0,531	0,657	0,156	0,367	0,841	0,576	0,458	0,323	0,063	0,098
Lpa*												
low	1,000			1,000			1,000			1,000		
medium	1,019	0,887	0,789	1,115	0,559	0,773	0,846	0,381	0,581	1,440	0,349	0,671
high	0,895	0,440	0,676	0,982	0,931	0,658	0,795	0,270	0,528	0,811	0,664	0,314
WHR**												
low	1,000			1,000			1,000			1,000		
medium	1,705	0,000	1,282	1,784	0,007	1,171	1,343	0,150	0,899	1,574	0,268	0,705
high	1,485	0,009	1,104	1,779	0,008	1,161	1,171	0,463	0,768	0,831	0,693	0,331
Ph.Cont.												
low	1,000			1,000			1,000			1,000		
medium	1,139	0,309	0,886	1,118	0,538	0,784	1,266	0,210	0,875	2,082	0,053	0,991
high	1,008	0,957	0,763	0,986	0,945	0,662	0,992	0,970	0,653	0,767	0,590	0,292
Smoke												
no	1,000			1,000			1,000			1,000		
former	1,017	0,887	0,807	1,054	0,755	0,759	0,981	0,913	0,695	0,973	0,940	0,476
current	1,406	0,130	0,905	1,662	0,083	0,937	1,275	0,495	0,635	2,474	0,115	0,802
Diabetes												
no	1,000			1,000			1,000			1,000		
yes	0,952	0,859	0,555	0,790	0,585	0,340	1,110	0,783	0,528	2,355	0,138	0,760
Alcohol												
low	1,000			1,000			1,000			1,000		
medium	0,910	0,584	0,651	1,205	0,492	0,708	0,747	0,216	0,470	0,768	0,614	0,275
high	0,845	0,315	0,609	1,056	0,838	0,626	0,722	0,157	0,460	1,200	0,709	0,461

Abbreviations: Opa= Occupational physical activity; Lpa= Leisure physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals.

Ph.Cont.= Physician contact.

^a H.R. are age-adjusted *lpa. low (<1.5 METh/day) medium (>1.5 <3.3 METh/day) high (>3.3 METh/day)

**waist to hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

Table 7.8: Associations between occupational and leisure physical activity and prostate cancer incidence/mortality, high education level

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality		
	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]
Oppa	1,000			1,000			1,000			1,000		
low	0,772	0,240	1,189	0,786	0,404	1,383	0,792	0,525	1,626	2,154	0,480	18,096
medium	0,349	0,033	0,919	0,450	0,209	1,563	0,337	0,171	1,600	0,000		
high												
Lpa*	1,000			1,000			1,000			1,000		
low	1,470	0,114	2,370	0,912	0,773	1,704	2,177	0,060	4,894	1,476	0,651	7,967
medium	1,526	0,107	2,552	1,343	0,363	2,537	1,521	0,365	3,775	3,931	0,093	19,458
high												
WHR**	1,000			1,000			1,000			1,000		
low	0,870	0,532	1,345	0,988	0,970	1,810	0,650	0,211	1,278	0,678	0,574	2,628
medium	0,645	0,073	1,042	0,742	0,371	1,427	0,509	0,081	1,087	1,077	0,911	3,964
high												
Ph.Cont.	1,000			1,000			1,000			1,000		
low	1,611	0,029	2,470	1,434	0,236	2,603	1,699	0,112	3,264	6,566	0,024	33,551
medium	1,360	0,205	2,186	1,544	0,179	2,910	1,060	0,884	2,315	4,944	0,057	25,648
high												
Smoke	1,380	0,107	2,042	1,362	0,247	2,300	1,452	0,244	2,719	1,253	0,719	4,281
no	1,385	0,494	3,521	0,771	0,726	3,292	1,989	0,362	8,724	0,000		
former												
current												
Diabetes	1,059	0,914	2,960	1,900	0,236	5,488	0,000	1,000	0,000	2,748	0,368	24,814
no												
yes												
Alcohol	1,000			1,000			1,000			1,000		
low	1,044	0,904	2,091	0,792	0,666	2,279	1,019	0,970	2,650	1,998	0,528	17,156
medium	1,122	0,729	2,152	1,401	0,485	3,613	0,783	0,604	1,970	0,762	0,806	6,660
high												

Abbreviations: Opa= Occupational physical activity; Lpa= Leisure physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals.
 Ph.cont.= Physician contact.
^aH.R. are age-adjusted *lpa. low (<1.5 METh/day) medium (>1.5 <3.3 METh/day) high (>3.3 METh/day)
 **waist to hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

Table 7.9: Associations between occupational (muscular activity) and leisure physical activity and prostate cancer incidence/mortality, low education level

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality			
	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	
Oppa (mus)													
low	1,000			1,000			1,000			1,000			
high	0,645	0,030	0,435	0,497	0,025	0,269	0,918	0,783	0,386	1,362	0,696	0,431	3,523
Lpa*													
low	1,000			1,000				1,000					
medium	1,013	0,924	0,785	1,108	0,583	0,769	1,597	0,837	0,350	1,216	0,449	0,630	2,841
high	0,883	0,382	0,667	0,958	0,833	0,642	1,429	0,790	0,256	1,186	0,529	0,291	1,885
WHR**													
low	1,000			1,000				1,000					
medium	1,704	0,000	1,281	1,782	0,007	1,169	2,715	1,343	0,149	2,007	0,241	0,725	3,603
high	1,481	0,009	1,101	1,780	0,008	1,161	2,727	1,167	0,474	1,779	0,745	0,343	2,147
Ph.Cont.													
low	1,000			1,000				1,000					
medium	1,135	0,324	0,883	1,117	0,541	0,784	1,592	1,260	0,220	1,823	0,047	1,009	4,462
high	1,002	0,991	0,758	0,978	0,911	0,656	1,457	0,989	0,959	1,503	0,619	0,297	2,060
Smoke													
no	1,000			1,000				1,000					
former	1,019	0,876	0,808	1,060	0,727	0,764	1,471	0,982	0,918	1,387	0,944	0,503	2,093
current	1,411	0,125	0,908	1,675	0,078	0,944	2,971	1,277	0,492	2,566	0,067	0,929	8,579
Diabetes													
no	1,000			1,000				1,000					
yes	0,957	0,874	0,559	0,798	0,600	0,343	1,855	1,117	0,771	2,345	0,096	0,845	8,058
Alcohol													
low	1,000			1,000				1,000					
medium	0,911	0,584	0,651	1,212	0,479	0,712	2,062	0,746	0,215	1,185	0,632	0,279	2,173
high	0,848	0,324	0,611	1,066	0,810	0,632	1,798	0,723	0,159	1,135	0,662	0,476	3,215

Abbreviations: Oppa= Occupational physical activity; Lpa= Leisure physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals.
 Ph.cont.= Physician contact.
^aH.R. are age-adjusted *lpa. low (<1.5 METH/day) medium (>1.5 <3.3 METH/day) high (>3.3 METH/day)
 **waist to hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

Table 7.10: Associations between occupational (muscular activity) and leisure physical activity and prostate cancer incidence/mortality, high education level

	Incidence (total)			Incidence (low-risk)			Incidence (high-risk)			Mortality		
	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]	H.R. ^a	P.V.	[95% C.I.]
Oppa (mus)												
low	1,000			1,000			1,000			1,000		
high	0,392	0,113	1,248	0,569	0,437	2,358	0,305	0,244	2,251	0,000		
Lpa*												
low	1,000			1,000			1,000			1,000		
medium	1,429	0,142	0,887	0,894	0,724	1,668	2,103	0,071	4,711	1,566	0,600	8,382
high	1,440	0,162	0,864	1,285	0,437	2,417	1,431	0,435	3,518	4,151	0,078	20,20
WHR**												
low	1,000			1,000			1,000			1,000		
medium	0,871	0,534	0,564	0,985	0,962	1,803	0,650	0,212	1,278	0,657	0,540	2,525
high	0,649	0,077	0,402	0,750	0,390	1,444	0,510	0,082	1,089	1,079	0,910	4,005
Ph.Cont.												
low	1,000			1,000			1,000			1,000		
medium	1,618	0,027	1,055	1,439	0,232	2,612	1,713	0,107	3,294	7,254	0,018	37,47
high	1,371	0,194	0,852	1,546	0,178	2,917	1,070	0,865	2,338	5,251	0,049	27,30
Smoke												
no	1,000			1,000			1,000			1,000		
former	1,394	0,097	0,942	1,377	0,232	2,326	1,466	0,233	2,746	1,203	0,771	4,187
current	1,426	0,455	0,561	0,795	0,756	3,389	2,034	0,346	8,920	0,000		
Diabetes												
no	1,000			1,000			1,000			1,000		
yes	1,028	0,958	0,368	1,847	0,258	5,341	0,000			2,822	0,351	25,01
Alcohol												
low	1,000			1,000			1,000			1,000		
medium	1,033	0,926	0,516	0,781	0,646	2,245	1,017	0,973	2,646	2,069	0,508	17,82
high	1,131	0,711	0,589	1,411	0,477	3,638	0,789	0,615	1,987	0,780	0,823	6,891

Abbreviations: Oppa= Occupational physical activity; Lpa= Leisure physical activity; H.R.=hazard ratio; P.V.=p-value; C.I.=confidence intervals.
 Ph.cont.= Physician contact.
^aH.R. are age-adjusted *lpa. low (<1.5 METH/day) medium (>1.5 <3.3 METH/day) high (>3.3 METH/day)
 **waist to hip ratio: low (<0.85) medium (>0.85 <0.90) high (>0.90)

Mediation analysis

Using mediation analysis, we were able to partition the total effect of physical activity on prostate cancer incidence into natural direct and indirect effects mediated through waist-to-hip ratio. Since strong and significant associations were found only for occupational muscular activity in relation to total and low-risk prostate cancer incidence, we performed a mediation analysis to test if this effect could have been mediated by waist-to-hip ratio decrease. No interactions between occupational muscular activity and waist-to-hip ratio was found. For this reason, indirect effect reduces to the product of parameters $\gamma_2\beta_1$ of models 4.15 and 4.18. The direct effect is represented by parameter γ_1 in model 4.18. When computing these quantities for total prostate cancer incidence, we then obtain a direct effect equal to 0.62 (C.I. 0.42-0.89) and an indirect effect equal to 0.99 (C.I. 0.98-1.01). When computing the same quantities for low-risk prostate cancer incidence we then obtain a direct effect equal to 0.51 (C.I. 0.29-0.90) and an indirect effect equal to 0.99 (C.I. 0.98-1.00).

7.5 Discussion

In this large prospective observational study, we found that total physical activity was not associated with prostate cancer incidence and mortality: physical active men had a comparable risk to develop or die from prostate cancer compared to sedentary men.

While physical activity did not seem to be involved in prostatic carcinogenesis and progression, a high waist-to-hip ratio represented a risk factor for low-risk prostate cancer incidence and mortality.

Our results on physical activity are consistent with 35% of findings of studies published on this topic. Two main reviews on the relation between physical activity and prostate cancer have been published: the first was published in 2001 by Friedenreich and Thune [37] and the second one was published in 2011 by Young [38].

Summarizing results from these two reviews, 55% of the studies highlighted a

protective role of physical activity for prostate cancer incidence, 10% had found an increased risk of prostate cancer associated with physical activity and the remaining had not shown any significant associations. Studies showing inverse associations, often found small effects. Of these studies, some considered occupational and leisure physical activity separately. While only a minority of studies highlighted a protective role of occupational physical activity, more evidence was found for vigorous leisure physical activity performed during young ages in terms of aggressive prostate cancer [38].

Our results showed a relation between highly demanding occupational physical activity and low-risk prostate cancer: when stratifying on educational level the association is no more significant, but association between heavy muscular activity and low-risk prostate cancer incidence remains significant.

One possible explanation for this result could be residual confounding due to socio-economical status. Low-risk prostate cancers are increasingly being discovered through screening procedures. As it has been shown for other types of cancers, screening coverage is not the same across different social classes [47]: our data showed that men with higher muscular activity were less educated, on average, than men with more sedentary jobs. Then, we could hypothesize that the protective effect of muscular activity on low-risk cancer to be only a consequence of less effective screening. Mediation analysis was performed to deepen this result: our findings show that waist-to-hip ratio cannot be considered a mediator in the analyzed relationship and we observe that all the protective effect of muscular activity on prostate cancer goes through pathways that involve other variables.

Contrary to what other studies have reported, leisure physical activity performed during young ages did not decrease risk of either low and high-risk prostate cancer.

We found that a high waist-to-hip ratio was a risk factor for total prostate cancer. This result is consistent with the findings of others [40].

Results present in literature are however not consistent and other studies found no statistically significant relationships between waist-to-hip ratio and risk of

prostate cancer [40]. A separate analysis by low and high-risk cancer showed that the relation found for total cancer holds only for low-risk tumors. It is worth noting that only the second tertile was significantly associated with higher risk, but not the third one. We could hypothesize that this result is a consequence of competing events, high waist-to-hip ratio being associated with overall mortality in men [48].

Similar biological mechanisms leading to prostate cancer have been suggested as a consequence of both waist-to-hip ratio increase and low physical exercise. For this reason, we tested if interactions between waist-to-hip ratio and physical activity (total, occupational and leisure) were supported by the data, but we did not find any effect modification. Thus, high levels of physical activity were not protective for prostate cancer, regardless of different amount of visceral fat. To our knowledge, few studies have investigated possible interaction between waist-to-hip ratio and physical activity. Indeed, more studies had focused on BMI and possible interactions with physical activity: we chose to consider in our analysis waist-to-hip ratio instead of BMI, because the former is a more specific measure of abdominal obesity which has been found to be in prostate carcinogenesis.

Many of the studies analyzing relationships between physical activity/waist-to-hip ratio and prostate cancer have shown that tumor type and grade at diagnosis can be affected in different way by hormonal and metabolic changes. For this reason, we performed analyses for both low and high-risk prostate cancer. Our results showed that only low-risk cancer was potentially related to physical activity and waist-to-hip ratio. No associations were found for high-risk cancer. In our data, 80% of fatal cancers are high-risk ones, suggesting that similar factors predict aggressive prostate cancer onset and progression to death.

We also performed separate analyses for mortality. Similarly to what was observed for high-risk cancer, physical activity was not associated with this outcome. Our results suggested that factors that trigger the disease are not the same that determine its progression.

The present study has several strengths. First, the data collection was conducted prospectively thereby eliminating bias from subject recall. The study design allowed for an assessment of the temporal sequence of the exposures and outcome. We used a validated instrument to assess total physical activity. We focused on either occupational or leisure time activity separately. Moreover, for occupational physical activity we did not just create an index to quantify it, but we also analyzed if the muscular or locomotive component had a major impact. In the same way, when considering leisure physical activity, we focused on different types of sport/outdoor activity, in order to analyze if type of exercise can have different impact on the studied outcomes.

Physical activity is a complex exposure to quantify and different aspects like duration, intensity and quality have all to be taken into account. To the best of our knowledge, our paper is one of the few ones that had the possibility to explore in such a detailed way different aspects of physical activity. This was possible because of the richness of the available data.

Due to the availability of healthcare and comprehensive data registers in Sweden, we also minimized misclassification of the outcome.

The population under study is homogeneous for ethnicity: it is well known that race has a role in prostate cancer incidence because of different underlying levels of testosterone among different populations [40]. In our study we did not have to adjust for this variable since more than 95% of the men declared that both parents' place of origin was Sweden.

The limitations of our study should be considered when interpreting the results. Given that both waist-to-hip ratio and physical activity (total, occupational and leisure) were self-reported, exposure may have been misclassified. In particular, contrary to what other studies have done, we chose not to use in the model BMI but waist-to-hip ratio: this choice us to a major risk of measurements errors and self reported waist-to-hip ratio measures should have been validated.

An other limitation of our study is represented by the assumption that we stated

for the mediation analysis: we hypothesized that there was a temporally shift between waist-to-hip ratio and physical activity measurements, since the latter was measured for the 12 months before the recruitment. This is however an assumption and longitudinal studies would be necessary in order to measure these two variables in different points in time.

The etiology of prostate cancer has been extensively discussed in literature, but still remains incompletely understood. Assessed risk factors for this cancer are increasing age, race and heredity. In our study we focused on potential modifiable risk factors, as physical activity and waist-to-hip ratio. More research is needed to understand the role of waist-to-hip ratio in the biological mechanisms and pathways that lead to prostate cancer and its possible interactions with other factors acting on hormones and metabolism. Mediation analyses considering multiple mediators could be useful and should be performed for a better clarification of these topics. Further research is also needed to identify physical activity patterns, in terms of quantity, intensity and type, that could be protective for prostate cancer incidence and progression.

Conclusions

The performed analyses allowed us to study the effect of different type of physical activity on prostate cancer incidence and mortality. We have not found any significant relationship of total and leisure physical activity with the outcomes, but we have found an effect of muscular occupational physical activity on prostate cancer incidence, when analyzing both total and low-risk prostate cancer. This result could be affected by residual confounding due to not complete adjustment for socio-economical level. However, the relationship remained significant also after stratifying for educational level, suggesting a potential causal effect. Possible future extensions of the work may include sensitivity analyses to assess potential bias due to social-economical status unmeasured confounding.

To understand better the pathway from muscular physical activity to prostate cancer, we analyzed the role of waist-to-hip ratio as a possible mediator. We have not found waist-to-hip ratio to be a mediator, meaning that the effect of heavy muscular activity on prostate cancer incidence goes through pathways that do not involve directly visceral fat reduction.

To clarify prostate cancer etiology it is thus necessary to focus more on pathways that involve other reasonable biological mechanisms. Physical activity is indeed known to increase the number of natural kills cells and to improve immune function by inducing changes in the activity also of neutrophils and macrophages. Since relations between immune functions and cancer etiology have been deeply studied, it is possible that physical activity could affect prostate cancer risk

through this path.

In the same way, physical activity could affect prostate cancer etiology by affecting free radical levels. It has been shown indeed that free radical production could be increased by acute physical exercise, but chronic one should improve free radical defenses activating free radical scavenger enzymes. Thus more research is needed to understand and quantify the role that these factors have in the pathway between physical activity and prostate cancer incidence.

We observed that waist-to-hip ratio does not moderate the effect of physical activity levels on prostate cancer risk. From a statistical point of view, this means that there is no significant interaction between the treatment and the mediator. The theory developed by VanderWeele reduces to the well known product method first proposed in literature by Baron and Kenny for continuous outcomes.

In this analysis, we have hypothesized a simple scenario where only one mediator was considered. However, there is growing recent literature on mediation models accounting for more than one mediator [49]. These designs have traditionally received less attention in both the methodological and applied literature even if researchers often have several putative mediators in mind to account for a given treatment-outcome relationship.

One of the main reason why there has been little focus on methods for testing multiple mediation hypotheses is probably that the analytic methods are not straightforward, relative to those for simple mediation and they are not available in standard routines in statistical softwares. To date, only a few authors [50, 51, 52, 53, 54] have devoted attention to the simultaneous testing of multiple indirect effects. To our knowledge, no author has handled this problem with an approach based on causal inference principles.

Thus, we suggest that methodological extensions of this work should involve study designs that consider multiple mediators simultaneously and we suggest

that methodologies formalized within the potential outcome framework should be developed. To the best of our knowledge, no author has dealt with these issues in survival analysis context yet. However, further research in this direction could be useful in all the situation in which a clarification of the biological mechanisms underlying a complex disease is needed.

Acknowledgements

First of all, I would like to thank my supervisor, Prof.Rino Bellocco who made possible the research work that is presented in this thesis. I express my gratitude to him for always suggesting interesting topics and ideas for my research. Thanks for having trusted and supported me throughout my whole Swedish experience, being constantly available for discussion and dialogue.

I thank all the members of the National March Cohort group, in particular Prof. Olof Nyren, Prof.Ylva Lagerros and Daniela Mariosa. I thank them for all the time spent together discussing my research results, for giving precious advice and for always encouraging. Thanks Daniela for being such an organized colleague!

Thanks to Arvid Sjolander for his availability and help as well as for inspiring me with his passion for causal inference.

I thank the Coordinator of the PhD program Prof.Fulvia Mecatti for being supportive and present during the years of doctoral studies.

My deep gratitude to Prof.Gianni Corrao, Prof.Antonella Zambon and Prof.Enzo Bagnardi for having introduced me to the world of epidemiology and for having always supported in these (almost 10!) years of university.

A huge thank to all my friends Vivi, Isa, Enrico, Silvia, Elisa, Marianna and, of course ...Sara!! Thanks for the invaluable help during these years and for encouraging me to take my chances (!) in far Sweden. Without you all, it would have been impossible to go through hard times and everything would have been much less fun!

Thanks to Ilia for the important experiences we shared since we met and for

everything we are still learning together.

Thanks Ago and Cori for distracting me from my numbers and helping me to be always in connection with myself.

Vale, Franci, Alice, Pier and Gabriele, you all contributed a lot to this thesis. Thanks, because I know I can rely on your sincere love in every moment of my life, wherever I am.

All this would not have been possible without the warmth and love of my family. Thanks to my mum, my dad and to my sisters Franci and Stefi. Thanks to my aunt and uncle, Bianca and Vitaliano, and to Chiara, Francesco and Elena. Your constant presence has made me feel warm even in the cold Sweden!

Bibliography

- [1] P.W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.
- [2] J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*, 5(4):465–480, 1923.
- [3] R.A. Fisher, S. Genetiker, R.A. Fisher, S. Genetician, G. Britain, R.A. Fisher, and S. Généticien. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh, 1970.
- [4] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.
- [5] D.B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591– 593, 1980.
- [6] P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- [7] R.B. D’Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19):2265–2281, 1998.
- [8] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

- [9] J. Pearl. Causality: models, reasoning, and inference. *IIE Transactions*, 34(6):583–589, 2002.
- [10] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, San Mateo, CA, 1988.
- [11] K.J. Rothman and S. Greenland. Causation and causal inference in epidemiology. *Journal Information*, 95(S1), 2005.
- [12] M. Hernan and J. Robins. *Causal inference*. Chapman & Hall/CRC, 2013.
- [13] R.M. Baron and D.A. Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.
- [14] C.M. Judd and D.A. Kenny. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5):602–619, 1981.
- [15] L.R. James and J.M. Brett. Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69(2):307, 1984.
- [16] K.J. Preacher, D.D. Rucker, and A.F. Hayes. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1):185–227, 2007.
- [17] D. Muller, C.M. Judd, and V.Y. Yzerbyt. When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89(6):852, 2005.
- [18] D.A. Scheufele. Examining differential gains from mass media and their implications for participatory behavior. *Communication Research*, 29(1):46–65, 2002.
- [19] D.P. MacKinnon and J.H. Dwyer. Estimating mediated effects in prevention studies. *Evaluation Review*, 17(2):144–158, 1993.

- [20] L.A. Aroian. The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, 18(2):265–271, 1947.
- [21] M.E. Sobel. Direct and indirect effects in linear structural equation models. *Sociological Methods & Research*, 16(1):155–176, 1987.
- [22] K.B. Karlson and A. Holm. Decomposing primary and secondary effects: A new decomposition method. *Research in Social Stratification and Mobility*, 29(2):221–237, 2011.
- [23] Psychometric Society. International Meeting. *New Developments in Psychometrics: Proceedings of the International Meeting of the Psychometric Society, IMPS2001, Osaka, Japan, July 15-19, 2001*. Springer Tokyo, 2003.
- [24] J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology Cambridge Mass*, 3(2):143–155, 1992.
- [25] C.E. Frangakis and D.B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.
- [26] K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309, 2010.
- [27] J. Pearl. Direct and indirect effects. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
- [28] T. J. Vanderweele and S. Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172(12):1339–1348, 2010.
- [29] T. Lange and J.V. Hansen. Direct and indirect effects in a survival context. *Epidemiology Cambridge Mass*, 22(4):575–581, 2011.
- [30] T.J. VanderWeele. Causal mediation analysis with survival data. *Epidemiology*, 22(4):582–585, 2011.

- [31] A. Heidenreich, J. Bellmunt, M. Bolla, S. Joniau, M. Mason, V. Matveev, N. Mottet, H.P. Schmid, T. van der Kwast, T. Wiegel, et al. Eau guidelines on prostate cancer. part 1: screening, diagnosis, and treatment of clinically localised disease. *European Urology*, 59(1):61–71, 2011.
- [32] O. Bratt et al. Hereditary prostate cancer: clinical aspects. *The Journal of Urology*, 168(3):906–913, 2002.
- [33] G.L. Andriole, E.D. Crawford, R.L. Grubb III, S.S. Buys, D. Chia, T.R. Church, M.N. Fouad, E.P. Gelmann, P.A. Kvale, D.J. Reding, et al. Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine*, 360(13):1310–1319, 2009.
- [34] F.H. Schröder, J. Hugosson, M.J. Roobol, T.L.J. Tammela, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa, et al. Screening and prostate-cancer mortality in a randomized european study. *New England Journal of Medicine*, 360(13):1320–1328, 2009.
- [35] E.M. Schaeffer, H.B. Carter, A. Kettermann, S. Loeb, L. Ferrucci, P. Landis, B.J. Trock, and E.J. Metter. Prostate specific antigen testing among the elderly when to stop? *The Journal of Urology*, 181(4):1606–1614, 2009.
- [36] S. Loeb and W.J. Catalona. What is the role of digital rectal examination in men undergoing serial screening of serum psa levels? *Nature Clinical Practice Urology*, 6(2):68–69, 2009.
- [37] C. Friedenreich and I. Thune. A review of physical activity and prostate cancer risk. *Cancer Causes & Control*, 12:461–475, 2001.
- [38] S. Young-McCaughan. Potential for prostate cancer prevention through physical activity. *World Journal of Urology*, 30:1–13, 2011.
- [39] E.A. Platz, M.F. Leitzmann, D.S. Michaud, W.C. Willett, and E. Giovannucci. Interrelation of energy intake, body size, and physical activity with prostate cancer in a large prospective cohort study. *Cancer Research*, 63(23):8542–8548, 2003.

- [40] B. Nemesure, S.Y. Wu, A. Hennis, and M.C. Leske. Central adiposity and prostate cancer in a black population. *Cancer Epidemiology Biomarkers & Prevention*, 21(5):851–858, 2012.
- [41] T. Pischon, H. Boeing, S. Weikert, N. Allen, T. Key, N.F. Johnsen, A. Tjønneland, M.T. Severinsen, K. Overvad, S. Rohrmann, et al. Body size and risk of prostate cancer in the european prospective investigation into cancer and nutrition. *Cancer Epidemiology Biomarkers & Prevention*, 17(11):3252–3261, 2008.
- [42] L.J. Su, L. Arab, S.E. Steck, E.T.H. Fontham, J.C. Schroeder, J.T. Bensen, and J.L. Mohler. Obesity and prostate cancer aggressiveness among african and caucasian americans in a population-based study. *Cancer Epidemiology Biomarkers & Prevention*, 20(5):844–853, 2011.
- [43] M.D. Jackson, S.P. Walker, C.M. Simpson, N. McFarlane-Anderson, F.I. Bennett, K.C.M. Coard, W.D. Aiken, T. Tulloch, T.J. Paul, and R.L. Wan. Body size and risk of prostate cancer in jamaican men. *Cancer Causes & Control*, 21(6):909–917, 2010.
- [44] J. Adolfsson, H. Garmo, E. Varenhorst, G. Ahlgren, C. Ahlstrand, O. Andrén, A. Bill-Axelsson, O. Bratt, J.E. Damber, K. Hellström, et al. Clinical characteristics and primary treatment of prostate cancer in sweden between 1996 and 2005. *Scandinavian Journal of Urology and Nephrology*, 41(6):456–477, 2007.
- [45] C.M. Friedenreich and M.R. Orenstein. Physical activity and cancer prevention: etiologic evidence and biological mechanisms. *The Journal of Nutrition*, 132(11):3456–3464, 2002.
- [46] Y.T. Lagerros, R. Bellocco, H.O. Adami, and O. Nyrén. Measures of physical activity and their correlates: the swedish national march cohort. *European Journal of Epidemiology*, 24(4):161–169, 2009.

- [47] E. Ward, A. Jemal, V. Cokkinides, G.K. Singh, C. Cardinez, A. Ghafoor, and M. Thun. Cancer disparities by race/ethnicity and socioeconomic status. *CA: A Cancer Journal for Clinicians*, 54(2):78–93, 2008.
- [48] P. Srikanthan, T.E. Seeman, and A.S. Karlamangla. Waist-hip-ratio as a predictor of all-cause mortality in high-functioning older adults. *Annals of Epidemiology*, 19(10):724–731, 2009.
- [49] K.J. Preacher and A.F. Hayes. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3):879–891, 2008.
- [50] K.A. Bollen. Total, direct, and indirect effects in structural equation models. *Sociological Methodology*, 17(1):37–69, 1987.
- [51] K.A. Bollen. *Structural Equations with Latent Variables*. Wiley, New York, 1989.
- [52] R.L. Brown. Assessing specific mediational effects in complex theoretical models. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(2):142–156, 1997.
- [53] M.W.L. Cheung. Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(2):227–246, 2007.
- [54] D.P. MacKinnon, J.L. Krull, and C.M. Lockwood. Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4):173–181, 2000.

