UNIVERSITEIT GENT

FACULTY OF SCIENCES

Mediation analysis of randomised experiments

Sjouke Vandenberghe

Proefschrift voorgedragen tot het behalen van de graad van
Doctor in de Statistische Data-Analyse
Academiejaar 2017-2018

Promotoren:
Prof. Dr. Stijn Vansteelandt
Prof. Dr. Luc Duchateau

Vakgroep Toegepaste Wiskunde, Informatica en Statistiek
Faculteit Wetenschappen, Universiteit Gent
Krijgslaan 281, B-9000 Gent

Table of Contents

Acknowledgements

Sjouke Vandenberghe
December 2017

CHAPTER 1

Introduction

The randomised controlled trial (RCT) is regarded as the gold standard for the evaluation of the causal effect of an exposure on an outcome. Next to primary intention-to-treat analyses, there is a growing recognition of the usefulness of mediation analysis in the analysis of randomised experiments. The goal of mediation analysis is to investigate the mechanisms that underlie the observed causal relationship between treatment and outcome: ‘opening the black box’ as Hafeman and Schwartz (2009) would say. In randomised controlled trials, mediation analyses can be used to identify why and how the treatment achieves its effect on the outcome by decomposing the total intention-to-treat effect into an *indirect effect* acting through a given intermediate variable, the so-called mediator, and the remaining *direct effect* (Kraemer et al. 2002; Oakley et al. 2006). As such, a mediation analysis may reveal the causal pathways through which a treatment affects a certain outcome of interest. Knowledge about these causal pathways may lead to a better understanding of how the treatment works. By focussing on active treatment elements and separating them from those that are inessential, better, more efficient, less expensive and possibly safer treatment regimes with fewer side effects may be developed (Kraemer et al. 2002). Consider, for example, the LEADER trial (Marso et al. 2016), where the primary analysis aimed to examine the effect and safety of liraglutide versus a placebo in adults with type 2 diabetes on time to first major adverse cardiovascular event (MACE), defined as time from randomisation to cardiovascular death, non-fatal myocardial infarction, or non-fatal stroke, ‘whichever came first’. Rather unexpectedly, the results in Marso et al. (2016) showed cardiovascular benefits of liraglutide as compared to the placebo treatment. Beneficial side effects such as weight loss, lower glucose levels and improved circulation as a result of a higher heart rate and lower systolic blood pressure were reported. Rasmussen and Nordisk (2016) hypothesized that these beneficial side effects could be potential mediators of the protective effect of liraglutide on time to first MACE. In this setting, mediation analysis could thus be used to examine the importance of potential mediators reduction in HbA1c (blood glucose level), weight loss and improved circulation from baseline on the hypothetical pathway from liraglutide to time to first MACE.

Furthermore, mediation analyses can be used to evaluate the effectiveness of an existing treatment to deal with a problem or disease for which it was not originally developed. In the evaluation of antidepressants, for example, researchers have noticed that certain antidepressants not only reduce symptoms of depression, but also improve cognitive function, which is known to be associated with major depressive disorder (MDD) (McIntyre et al. 2013). The efficacy of vortioxetine (a rather novel antidepressant) on depressive symptoms, for instance, has already been demonstrated in several short-term studies (6 to 8 weeks) (Alvarez et al. 2012; Boulenger et al. 2014). Its effect on cognitive function was further examined in the randomised controlled trial of McIntyre et al. (2014), which showed a beneficial effect of vortioxetine on cognition and more importantly that this effect is mainly due to a direct treatment effect, and not because it eased depressive symptoms (i.e. indirect effect). Such results may convince the scientific community of the usefulness of vortioxetine in improving cognitive function, even in the absence of depressive symptoms, for example to improve residual cognitive function deficits in remitted patients.

In randomised trials, mediation analyses are also recommended when the interpretation of the intention-to-treat effect is hindered by the presence of intermediate variables. One such example is the MIRA trial (Padian et al. 2007; Rosenblum et al. 2009), a randomised controlled trial in HIV-negative, sexually active women in South Africa and Zimbabwe that aimed to asses the effect of diaphragm and lubricant gel use in reducing HIV infection. The 5045 female participants were randomised to either the placebo treatment consisting of intensive condom counselling and provision (i.e. the usual HIV prevention method) or the active treatment that additionally provided women with a latex diaphragm and lubricant gel. Unfortunately, the results showed that the proportion of women using condoms was significantly lower in the active treatment than in the placebo group, which made it difficult to draw conclusions about diaphragm and gel efficacy. With a mediation analysis, Rosenblum et al. (2009) isolated the effect of diaphragm and lubricant gel on HIV infection mediated by condom use from the remaining direct effect and thereby provided more information about diaphragm and gel efficacy had condom use not been influenced, which is considered an important public health question. Similar complications arise in randomised trials that include the possibility of rescue medication, if this occurs more frequently in one of the treatment arms. The intention-to-treat effect, for instance, will not fully reflect the potential beneficial effect of an active treatment on survival if patients that received the placebo treatment more frequently receive additional rescue medication.

Finally, mediation analyses in randomised trials are also useful to examine mediating variables that are possible indicators of later disease, so-called *intermediate endpoints* or *surrogate markers*. Consider the EORTC 10994/BIG 1-00 randomised phase 3 trial (Bonnefoi et al. 2011), where women with locally advanced inflammatory or large operable breast cancer were randomly assigned to one of two neoadjuvant treatments before undergoing primary surgery. The study was set up to examine whether TP53 status could be used to predict if women would be more sensitive to anthracycline or taxane based neoadjuvant chemotherapies. At the time of surgery, pathological complete response (pCR), i.e. complete disappearance of any invasive cancer in the primary tumour and lymph nodes with the exception of very few scattered tumour cells, was measured. In neoadjuvant trials, pCR is often used as an endpoint to support accelerated approval of a drug for high-risk, early stage breast cancer, such that patients can be provided (conditional) access to promising drugs while confirmatory clinical trials are being conducted. Mediation analysis methods can be used to examine the validity of intermediate endpoints such as pCR. Identifying these surrogate markers is useful because replacing rare or late-occurring clinically relevant outcomes by frequent or short-term outcomes, might substantially reduce the cost and duration of experiments. But although the benefits are obvious, caution is in order. Inappropriately using a surrogate endpoint instead of the outcome of interest may give misleading results which in turn may lead to improper treatment for a large group of patients. One can find examples in studies about cardiovascular disease and colorectal cancer where arrhythmia and tumour response respectively are often used as an intermediate endpoint for overall survival. The Cardiac Arrhythmia Suppression Trial Investigators (1989) showed that the two major antiarrhythmic drugs encanaide and flecanaide indeed reduced arrhythmia initially, but also more than tripled the risk of eventually dying from arrhytmia. In advanced colorectal cancer, Buyse et al. (2000b) concluded, based on their meta-analysis of 25 randomised trials, that an increased tumour response was indeed related to an increase in overall survival. Knowing the effect of treatment on tumour response, however, seemed insufficient to predict its ultimate effect on overall survival. Using arrhythmia and tumour response as surrogate endpoints for overall survival in clinical trials might thus not be appropriate.

1.1 Mediation history

Mediation analysis is a popular topic nowadays, but the first attempt at quantifying mediating mechanisms in fact began in the early twentieth century with Wright's path analysis (Wright 1920). Wright's path analysis visually displays the variables and the relations between them in a path diagram and generates coefficients that represent the strength of the relationship between those variables. It was Wright who showed that the mediating process could be quantified as the product of all the path coefficients in the chain of mediation. Later on, structural equation modeling became popular when the path analysis approach was rediscovered by sociologists and economists (Duncan 1966; Simon 1954) and was made more general by combining it with covariance structure modeling (Bentler 1980). These structural equation models combine a structural model that shows dependencies between the variables and a measurement model showing relations between the latent variables and their indicators. As such, path analysis is a special case of structural equation modeling, consisting of only the structural part (i.e. no latent variables).

In the 1970s, researchers from several fields became aware of the usefulness of mediation analysis. In social sciences, structural equation modeling was revisited with two landmark papers about mediation: one from Judd and Kenny (1981), who used mediation analysis for the development and evaluation of disease prevention programs, and one from Baron and Kenny (1986), that described methods to examine mediation and moderation in detail. MacKinnon (2008) gives an overview of these statistical procedures typically used in social sciences to conduct mediation analysis for single and multiple mediators. In the medical and epidemiological literature, Prentice (1989) examined the existence of mediating variables as possible surrogate markers. He provided a formal definition and rather strict criteria to validate surrogate endpoints, because a variable could only be a surrogate marker if it fully captured the treatment effect on the primary endpoint. In practice, a surrogate endpoint is more likely to only explain part of the treatment effect. Therefore, Freedman et al. (1992) suggested the *proportion-explained approach*, better known as the Freedman approach, a quantitative measure of the proportion of the effect of the treatment on the outcome that is explained by the surrogate.

1.1.1 Traditional mediation analysis with a single mediator

Before providing a detailed description of modern causal mediation analysis approaches, we first discuss the traditional approach and the problems accompanying it. We will make use of graphs, which we will refer to as causal diagrams, to visualize this. Figure 1 shows the causal diagram of an ideal randomised study (i.e., no loss to follow-up, full treatment adherence, and double blind assignment) with treatment A and end-of-study outcome Y.

In deze sectie
moest ik letten
op het verschil tussen
schatters en echte
waarden, is dit nu beter?

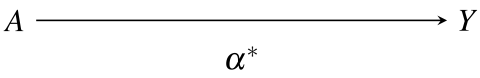


Figure 1: Causal diagram of the effect of treatment A on outcome Y.

For a continuous outcome Y, the intention-to-treat effect can be estimated via a linear regression model of Y on A:

$$\mathbb{E}(Y|A) = \mu_1 + \alpha^* A, \tag{1.1}$$

where α^* represents the intention-to-treat effect, for which an unbiased (in large samples) estimate can be obtained by substituting α^* by an ordinary least squares estimate.

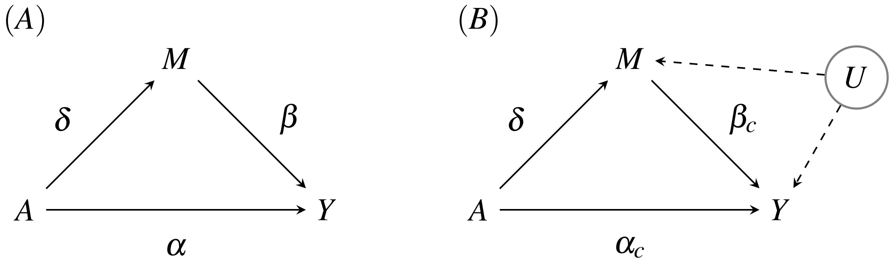


Figure 2: Causal diagram of randomised treatment A, mediator M, outcome Y (A) and confounders U (B).

Figure 2a again illustrates a causal diagram of an ideal randomised experiment with treatment A, mediator M and continuous outcome Y. Two additional linear regression models are used to investigate mediation:

$$\mathbb{E}(Y|A,M) = \mu_2 + \alpha A + \beta M \tag{1.2}$$

and

$$\mathbb{E}(M|A) = \mu_3 + \delta A. \tag{1.3}$$

The direct effect of A on Y not going through intermediate variable M, α is estimated via the regression of Y on A with adjustment for M. Even if A is randomised, one can not be sure that $\hat{\alpha}$, an estimator of α, results in an unbiased estimate of the direct treatment effect on outcome Y. The problem is that while A is randomised, the mediator M is not. And thus, it is very likely that there are unmeasured common causes U of mediator M and outcome Y (Figure 2b) and even if they would have been measured, the linear regression model in (1.2) does not control for them. As a result, $\hat{\alpha}$ is not an unbiased estimator because the estimate will not only contain the direct effect $A \rightarrow Y$, but also a spurious association via $A \rightarrow M \leftarrow U \rightarrow Y$. To gain insight, imagine a causal diagram with randomised treatment A (e.g., antidepressants), mediator M (e.g., depressive symptoms), outcome Y (e.g., cognitive function) and common cause U (e.g., physical activity level). Although the treatment and one's physical activity level are not associated by themselves, this changes if we condition on depressive symptoms. One can easily imagine that those patients who experience a decrease in their depressive symptoms, but were not physically active, were more likely the ones that received antidepressants and not a placebo. As such, the regression coefficient α would not only consist of the direct effect of antidepressants on cognitive function (i.e., α_c), but also of an additional source of association due to the inclusion of depressive symptoms into the regression model, because this induces an association between treatment and physical activity level, which is also associated with cognitive function.

The effect of treatment A on outcome Y that goes through the intermediate variable M can be estimated via two different approaches. We can see this if we use equations (1.2) and (1.3) to rewrite $E(Y|A)$ from (1.1) as follows:

$$\begin{aligned}
\mathbb{E}(Y|A) &= \mathbb{E}\{\mathrm{E}(Y|A,M)|A\} \\
&= \mathbb{E}\{\mu_2 + \alpha A + \beta M|A\} \\
&= \mu_2 + \alpha A + \beta \mathbb{E}\{M|A\} \\
&= \mu_2 + \alpha A + \beta(\mu_3 + \delta A) \\
&= \mu_2 + \beta\mu_3 + (\alpha + \beta\delta)A \qquad (1.4)
\end{aligned}$$

Then μ_1 and $\alpha^* A$ from (1.1) equal $\mu_2 + \beta\mu_3$ and $(\alpha + \beta\delta)A$ from (1.4) respectively. The latter equality can be rewritten as:

$$\alpha^* - \alpha = \delta\beta \qquad (1.5)$$

which shows the two approaches to calculate the indirect effect. Estimates for the indirect effect can be obtained by substituting α^*, α, δ and β by ordinary least squares estimates. Substituting α^* and α on the left-hand side, results in an indirect effect estimated via the *difference-of-coefficients* method, where the indirect effect equals the difference between the total and the direct effect. Substituting δ and β on the right-hand side equals the mediated or indirect effect according to the *product-of-coefficients* method. The total intention-to-treat effect, α^*, can thus be decomposed into a direct effect, α, and an indirect effect $\delta\beta = \alpha^* - \alpha$. Freedman et al. (1992) similarly used the difference between α^* and α as a measure of the indirect effect and additionally divided this by the total intention-to-treat effect α^*, which results in the proportion of the effect of treatment A on outcome Y that is explained by mediator M, the so-called *proportion mediated*.

Similar to the problem with $\hat{\alpha}$, an estimator of α in equation (1.2), $\hat{\beta}$ will not be an unbiased estimator of β (i.e. the effect of mediator M on outcome Y), because M is not randomised. Unlike $\hat{\alpha}^*$ estimator of α^* in equation (1.1) and $\hat{\delta}$ estimator of δ in equation (1.3) that are unbiased because treatment A is randomised, we can not be sure about $\hat{\alpha}$ and $\hat{\beta}$ if we have not measured all common causes of M and Y or do not include them in our analysis. This bias, induced because common causes of the mediator and outcome are not included, is called *confounding bias*. Although Judd and Kenny (1981) and MacKinnon (2008) recognise the need to control for important confounders of the mediator - outcome relationship, even in randomised trials, this was not pointed out by Baron and Kenny (1986) and thus frequently ignored in the social sciences literature.

A main concern of this traditional approach is that it was originally developed for linear regression models. Assuming linear relationships is no longer appropriate however, when the outcome is not measured at the interval level. As a result, the method was extended to non-linear models and is routinely used in those settings, but it has no formal justification and the interpretation of the effect measures is not well-defined (Lin et al. 1997; Kaufman et al. 2004; Imai et al. 2010; Pearl 2012). Tein and MacKinnon (2003) apply this approach to time-to-event outcomes, with accelerated failure time and proportional hazard models, but their analysis has severe shortcomings. They use model (1.3) for the mediator and two proportional hazard models for the time-to-event outcome

$$\lambda(t) = \lambda_0(t)\exp\{\gamma A + \eta M\} \tag{1.6}$$

and

$$\lambda(t) = \lambda_0(t)\exp\{\gamma^* A\} \tag{1.7}$$

or the log hazard form of the models

$$\log\{\lambda(t)\} = \log\{\lambda_0(t)\} + \gamma A + \eta M \tag{1.8}$$

and

$$\log\{\lambda(t)\} = \log\{\lambda_0(t)\} + \gamma^* A. \tag{1.9}$$

and assume that there is no censoring. If the population regression parameters γ^*, γ, η and δ are substituted with their appropriate estimators $\hat{\gamma}^*$, $\hat{\gamma}$, $\hat{\eta}$ and $\hat{\delta}$, then Tein and MacKinnon (2003) show via simulations that with proportional hazards models the indirect effect estimated via the difference-of-coefficients method (i.e. $\hat{\gamma}^* - \gamma$) differs from the product-of-coefficients estimate of the indirect effect (i.e. $\hat{\eta}\hat{\delta}$). VanderWeele (2011) shows, provided that the outcome is rare and there are no treatment-mediator interactions, that the difference-of-coefficients method and product-of-coefficients method approximately coincide. In general, however, neither the product-of-coefficients or the difference-of-coefficients method for the proportional hazards model have a clear causal interpretation as a measure of effect.

Additionally, VanderWeele (2011) shows that even if the outcome is not rare, the product-of-coefficients method does provide a valid test for whether there is an indirect effect of treatment on the outcome, provided the models are correctly specified and that the assumptions for natural direct and indirect effects, discussed later on, hold. That this is true can be understood as follows: $\hat{\eta}\hat{\delta}$ will only be different from zero if both $\hat{\eta}$ and $\hat{\delta}$ are different from zero. If the assumptions for natural direct and indirect effects hold and the models for the outcome and mediator are correctly specified then from $\hat{\delta} \neq 0$ it follows that A has a direct effect on M and from $\hat{\eta} \neq 0$ it follows that M has a direct effect on Y and as a result that the natural indirect effect is non-zero. The difference-of-coefficients method, on the other hand, can potentially result in an effect estimate for the indirect effect that is different from zero in non-linear models even when there is no indirect effect. This problem of non-linear models is called **non-collapsibility** (Greenland et al. 1999). Imagine a simple example (Figure 3) with a binary randomised exposure A, a covariate M and a time-to-event outcome Y.

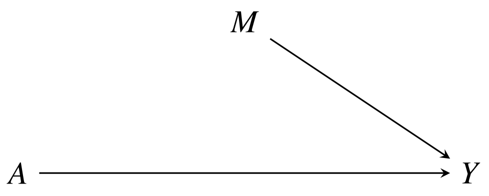


Figure 3: An example of non-collapsibility.

Fitting a Cox regression model with and without adjustment for M may result in estimates of a systematically different magnitude for the intention-to-treat effect, even when M is not a mediator, as in the example, because exposure A has no effect on M (Martinussen and Vansteelandt 2013). Another concern with the difference-of-coefficients method is **model incongeniality**. If there are two Cox regression models for outcome Y that are fitted simultaneously, then in general they are unlikely to be true at the same time (Lin et al. 1997; Bycott and Taylor 1998).

In the surrogacy setting, Freedman's approach received the criticism that the proportion explained is unstable in small samples or for small effect sizes (Daniels and Hughes 1997; Lin et al. 1997; Buyse and Molenberghs 1998; Bycott and Taylor 1998; Freedman 2001; MacKinnon et al. 1995). It only seemed useful in situations with a highly significant total treatment effect or large sample sizes. Otherwise this proportion explained could possibly lie outside the $[0,1]$ range and the confidence intervals frequently cover the whole $[0,1]$ interval, which is too wide to be useful.

1.1.2 Traditional mediation analysis with multiple mediators

Often more than one mediational process is of interest in the study of the relationship between the treatment and outcome of interest. The Multiple Risk Factor Intervention Trial (MRFIT) Group (1990), for instance, designed the trial to study the effect of treatment on prevention of heart disease via three possible mediators: smoking, cholesterol and blood pressure. At other times, a scenario with multiple mediators presents oneself when the relationship between the mediator and outcome of interest is confounded by a second mediator. In the Multiple Risk Factor Intervention Trial, for instance, there is a good chance that smoking is both mediator and, since it influences cholesterol and blood pressure, a confounder of the relationship between the outcome and the other mediators. The traditional literature on structural equation models (MacKinnon 2008) provides a framework to deal with multiple mediators that is a straightforward extension of their single mediator approach.

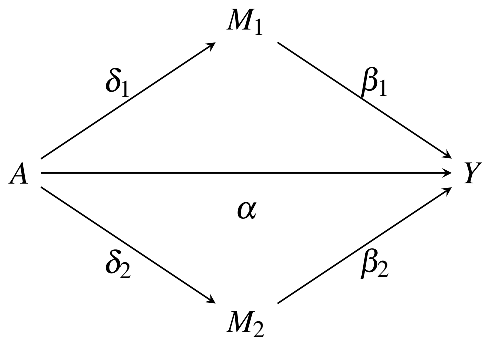


Figure 4: Causal diagram of randomised treatment A, mediators M_1 and M_2, and outcome Y.

Four regression equations are used to study mediation in the simple two-mediator model represented in Figure 4.

$$\mathbb{E}(Y|A) = \mu_1 + \alpha^* A, \tag{1.10}$$

$$\mathbb{E}(Y|A, M_1, M_2) = \mu_2 + \alpha A + \beta_1 M_1 + \beta_2 M_2 \tag{1.11}$$

$$\mathbb{E}(M_1|A) = \mu_3 + \delta_1 A. \tag{1.12}$$

$$\mathbb{E}(M_2|A) = \mu_4 + \delta_2 A. \tag{1.13}$$

In the traditional mediation analysis approach edges in sequence can be multiplied to obtain effects of interest. As such, the product of parameters δ_1 and β_1, i.e. $\delta_1\beta_1$, and the product of parameters δ_2 and β_2, i.e. $\delta_2\beta_2$, are assumed to represent the mediated or indirect effects via M_1 and M_2 respectively. Similarly to the single mediator model, the direct effect of A on Y not going through any intermediate variable, α is estimated via the regression of Y on A with adjustment for M_1 and M_2. This means that the total intention-to-treat effect, α^*, can be decomposed into a direct effect, α, and a total indirect effect $\alpha^* - \alpha$ or $\delta_1\beta_1 + \delta_2\beta_2$ (i.e. the difference-of-coefficients and the product-of-coefficients method respectively, when substituted with their respective estimators). The latter follows from the fact that parallel effects (i.e., $A \to M_1 \to Y$ and $A \to M_2 \to Y$) can be summed to obtain the effect of interest.

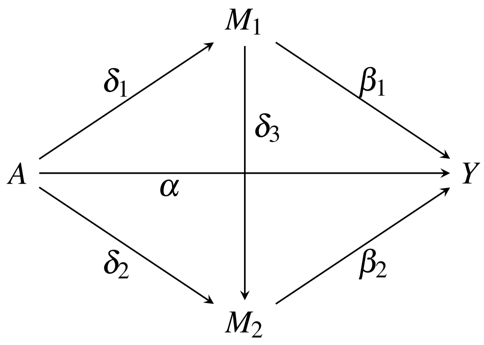


Figure 5: Causal diagram of randomised treatment A, sequential mediators M_1 and M_2, and outcome Y.

A slightly more complicated setting arrises when M_1 is not only a mediator, but also a confounder of the relationship between the outcome and the second mediator, as in Figure 5. In the regression equations, mediator M_1 is added to the regression equation for M_2, with δ_3 as parameter. As for the effects, similarly, the direct effect of A on Y not going through any intermediate variable equals α. Mediated effects are estimated via the product of the coefficients for each of the paths in the causal pathway and the sum of parallel pathways. Thus, the total indirect effect of A on Y is calculated as $\delta_1\delta_3\beta_2 + \delta_1\beta_1 + \delta_2\beta_2$. In the traditional mediation analysis approach, this effect can be broken down into three-paths: the effect passing via both mediators $\delta_1\delta_3\beta_2$, and the effects via only one of the mediators $\delta_1\beta_1$ and $\delta_2\beta_2$.

The structural equation model approach to traditional longitudinal mediation analyses of continuous and binary outcomes (MacKinnon 2008) and extensions to time-to-event outcomes, so-called dynamic path analysis (Fosen et al. 2006; Strohmaier et al. 2015) have received a lot of critique. Next to the fact that they deliver no meaningful or vague interpretations of the effects, these techniques were originally developed for linear regression analysis and have no justification in non-linear models (VanderWeele and Vansteelandt 2009, 2010; Imai et al. 2010). Outside of linear models the difference-of-coefficients method falls short and the product-of-coefficients method is only valid under very stringent parametric constraints (Taylor et al. 2008), such as the combination of specific parametric models, for example an additive hazard model for the time-to-event (due to its collapsibility properties) and linear regression models for the mediators in survival settings (Fosen et al. 2006; Strohmaier et al. 2015). Literature on the topic is also rather vague about their assumptions of unmeasured confounding: they logically assume no unmeasured treatment-outcome and mediator-outcome confounding, but also implicitly assume the absence of unmeasured confounders of the multiple mediators and measured time-varying confounders and the outcome. Additionally, they assume the absence of long term effects of covariates and mediators on covariates and mediators measured later in time. Further, they work under a strong no-interaction assumption on the individual level: the difference in outcome that would have been observed for a patient under active and placebo treatment with the mediator fixed to equal m is constant for all values of m (De Stavola et al. 2015). This assumption is biologically rather unlikely however and known to be violated in the presence of treatment - mediator interactions and with dichotomous and time-to-event endpoints (Robins and Greenland 1992; Robins 2003; De Stavola et al. 2015). Finally, they provide effect estimates for all path-specific effects in Figure 5. Later, it will be shown that such an effect decomposition is unrealistic as the indirect effect via M_1 and M_2 (i.e., $A \to M_1 \to M_2 \to Y$) can never be separated from the effect only going through M_1.

The extensive number of papers published in the last decade in medical and epidemiological journals on the topic of mediation analysis in randomised and non-randomised studies (Ananth and VanderWeele 2011; Lynch et al. 2008; Nandi et al. 2012; Oba et al. 2011; Rochon et al. 2014; Wang et al. 2010) illustrates that, recently, the scientific community indeed has taken a big interest in the analysis of direct and indirect effects of an exposure on an outcome. Extending the social sciences literature of the 1980s (Baron and Kenny 1986; Judd and Kenny 1981), methodological advances in the causal inference literature, which we will describe in the next section, have provided a way to formally decompose these effects under certain well-defined conditions.

1.2 Methodological advances in mediation analysis

1.2.1 Potential Outcomes

Stoort het in Secties 1.2.1, 1.2.2 en 1.2.3 dat er om specifieke counterfactuals duidelijk te maken, het uitgelegd wordt voor patient i terwijl de def. en assumpties algemeen zijn (zonder i's)?

First some potential outcome or counterfactual notation is introduced. This formal framework, used to clarify causal effects and conditions that enable estimating these effects, is actually very intuitive because people are used to reasoning in terms of hypothetical situations in everyday life: “If I take an aspirine now instead of doing nothing, my headache will be gone within 30 minutes”.

Let’s get back to the LEADER trial, ignoring censoring and the typical skewness of survival times for now. There the causal effect of liraglutide on time to first MACE can be defined as the difference between two potential outcomes: the time to first MACE that would be realised if the patient received liraglutide and the one that would be realised if the patient received the placebo. Let X denote the baseline covariates and possible confounders of the relationship between blood glucose, the mediator M and time to first MACE, the outcome Y. Further, let A denote the randomised treamtent with two treatment arms, the placebo $(A = 0)$ and the experimental treatment liraglutide $(A = 1)$, with known randomisation probabilities. Then, we can use $M(a)$ and $Y(a)$ to denote the potential blood glucose level and the potential time to first MACE that would occur under treatment status a. Thus, $M_i(0)$ and $M_i(1)$ measure individual i's blood glucose level if this individual would have been randomised to receive the placebo or liraglutide, respectively. Similarly, $Y_i(0)$ and $Y_i(1)$ measure what would have happened to the potential time to first MACE of individual i, if this individual had been randomised to receive the placebo or liraglutide, respectively. Even though there are two potential outcomes for each patient, only one of them is actually observed. For instance, if patient i was randomised to the placebo treatment, then only $Y_i(0)$ is observed and $Y_i(1)$ remains unobserved for that patient. The causal effect of liraglutide on time to first MACE for patient i can subsequently be defined as $Y_i(1) - Y_i(0)$. It is impossible to identify this unit-level causal effect, even in randomised experiments, because only one of these two counterfactual outcomes are observed. Therefore, the focus of researchers is on the estimation of the population-average causal effect defined as

$$\mathbb{E}\{Y(1) - Y(0)\}.$$

1.2.2 Natural Direct and Indirect Effects

But what if researchers from the LEADER trial (Rasmussen and Nordisk 2016) were interested in whether the protective effect of liraglutide on time to first MACE is due to its effect on blood glucose level. Robins and Greenland (1992) extended the counterfactual outcome notation to define causal mediation effects by introducing *nested counterfactuals*. Previously, the potential outcomes (e.g. $Y(a)$) were a function of the treatment, while in mediation analysis a potential outcome (e.g. $Y(a, M(a))$) is itself a function of the treatment and another potential outcome $M(a)$. Mediation analysis does not only rely on nested counterfactuals such as $Y(a, M(a))$, but also makes use of *cross-world counterfactuals* like $Y(a, M(a^*))$. These nested counterfactuals are referred to as cross-world counterfactuals because they require information from a single patient in two separate "worlds" a and a^*. Using these nested counterfactuals, we can now define the indirect effect via mediator M for each patient i as

$$IND_i(a) = Y_i(a, M_i(1)) - Y_i(a, M_i(0))$$

with $a = 0$ the *pure* indirect effect and $a = 1$ the *total* indirect effect (Robins, 2003). These indirect effects show what would have happend to the time to first MACE for patient i had this patient's blood glucose level been changed from $M_i(0)$, the blood glucose level he would have had, had he been randomised to the placebo arm, to $M_i(1)$, his potential blood glucose level on the liraglutide arm, without changing the actual treatment arm (i.e. holding it fixed at a). For $a = 1$ for example, $Y_i(1, M_i(1))$ equals the observed time to first MACE if this patient would have been randomised to the liraglutide treatment arm. Whereas $Y_i(1, M_i(0))$ represents what the time to first MACE for patient i would have been, had this patient been randomised to the liraglutide treatment, but with a blood glucose level equal to the one he would have had if he would have been randomised to the placebo arm. Note that the counterfactual $Y(1, M(0))$ is itself a function of another counterfactual $M(0)$ but with a different a status, as such, $Y(1, M(0))$ is unobservable, but it allows the formal expression of the indirect effect of treatment on outcome.

Similarly, the direct treatment effect for patient i can be defined as

$$DE_i(a) = Y_i(1, M_i(a)) - Y_i(0, M_i(a))$$

with $a = 0$ the *pure* direct effect and $a = 1$ the *total* direct effect (Robins, 2003). The direct effect of treatment is thus the change in time to first MACE had patient i been randomised to the liraglutide treatment instead of the placebo, while holding his blood glucose level fixed at the level it would have been naturally under treatment regime a. If we assume that the direct and indirect effects do not vary in function of treatment status (i.e., no difference in the pure and total direct and indirect effects: $IND = IND(1) = IND(0)$ and $DE = DE(1) = DE(0)$), then note that the sum of either one of these direct and indirect effects equals the total intention-to-treat effect. If the direct and indirect effects do differ in function of treatment status, then one total and one pure effect sum to the total intention-to-treat effect. In practice, researchers are typically interested in the average direct effect

$$DE(a) = \mathbb{E}\{Y(1, M(a)) - Y(0, M(a))\},$$

and the average indirect effect

$$IND(a) = \mathbb{E}\{Y(a, M(1)) - Y(a, M(0))\}$$

for $a = 0$ or 1. Pearl (2001) referred to these effects as *natural* direct and indirect effects, in contrast to *controlled* direct effects $E\{Y(1, m) - Y(0, m)\}$, because the natural direct effect captures the treatment effect upon fixing the mediator to the level it would have *naturally* taken for a patient under treatment regime a, rather than fixing it to be m for all patients. The strength of this potential outcome framework is that it defines causal direct and indirect effects without reference to specific statistical models. They can accommodate all kinds of linear and nonlinear relationships and various types of mediators and outcome variables irrespective of the scale of interest. Previously, we defined the effects in terms of mean differences, but they can be defined in terms of risk ratios and odds ratios as well.

Sceptics criticize the fact that natural direct and indirect effects are defined as a contrast of potential outcomes (i.e. outcomes that are not directly observed), but would have been observed if the mediator and/or exposure would have been different. Moreover, *cross-world counterfactuals* are involved, as they are potential outcomes that can never be observed in practice. Because scientists are generally interested in investigating the effect of a realistic intervention, this raises the question whether natural direct and indirect effects are all that informative if they do not even correspond to some realistic event. This does not mean however that natural direct and indirect effects should be forgotten altogether. First, they can be identified if one is willing to make specific assumptions. Also note that statisticians continuously seek for ways to relax these and make them more realistic. Secondly, many causal effects can not be estimated via randomised trials for ethical or practical reasons, but this should not be the reason why one should not try to get as close to the truth as possible (Naimi et al. 2014).

1.2.3 Identification

The nested counterfactuals that enable the formal expression of natural direct and indirect effects are related to the observed variables through the *consistency assumption*. If this assumption holds, then interventions that set exposure A to 1 (or 0) have no effect amongst those for whom the observed exposure level was $A = 1$ (or $A = 0$). This also holds for interventions on the mediator. In mediation analyses, we additionally make the *composition assumption* that $Y(1,M(1)) = Y(1) = Y$ if $A = 1$. This implies that mediation analyses give us the results of non-invasive interventions or manipulations (VanderWeele and Vansteelandt 2009), thus, changing the exposure while holding the mediator at a fixed, but individual-specific in case of natural effects, level.

Natural direct and indirect effects conceptualize an intervention on both the mediator and the exposure and their identification requires specific assumptions besides the consistency and composition assumption. Because they conceptualize an intervention on the exposure, one has to control for all confounders of both the exposure-mediator and the exposure-outcome relationship in the sense that

$$\forall\, a,m:\ Y(a,m) \perp\!\!\!\perp A|X, \tag{1.14}$$

and

$$\forall\, a:\ M(a) \perp\!\!\!\perp A|X. \tag{1.15}$$

Thus, the counterfactual outcome $Y(a,m)$ and the counterfactual mediator $M(a)$ should be independent of the actual treatment A within strata of a set of covariates X. This is the so-called *conditional ignorability* assumption, which states, for example in (1.14) with Y mortality, that within strata of X the risk of dying that would have been observed under treatment regime $a = 1$ and mediator level m, would be the same in the treatment group that actually got $A = 1$ as in the $A = 0$ group if they would have received, contrary to the fact, the same treatment as the $A = 1$ group. Also called the *no unmeasured confounding* assumption, these assumptions show that X has to be a sufficient set of covariates to adjust for possible confounding of the exposure-outcome and exposure-mediator relationship. Since we focus on randomised controlled trials, we do not need to worry about these two assumptions, because randomisation is expected to produce conditional ignorability with respect to the randomised treatment A. But randomisation in itself is not enough. The mediator is not randomly assigned and identification of natural direct and indirect effects, even in randomised trials, demands additional control for confounding of the mediator-outcome relationship. One thus has to assume that all confounders X of the mediator-outcome association have been measured in the sense that

$$\forall\, a,m: \; Y(a,m) \perp\!\!\!\perp M|A=a,X. \tag{1.16}$$

This states that the counterfactual outcome $Y(a,m)$ that, possibly contrary to the fact, would have been observed under an intervention that sets $A=a$ and $M=m$, does not depend on the actual level M within strata with $A=a$ and a set of covariates X. Not only do we need information about nested counterfactuals such as $Y(1,M(1))$ in the subjects for whom $A=1$ is actually observed, we also need to learn about the distribution of cross-world counterfactuals such as $Y(1,M(0))$. Pearl (2001) assumes that

$$\forall\, a,a^*,m: \; Y(a,m) \perp\!\!\!\perp M(a^*)|X. \tag{1.17}$$

If we assume that the data are generated under a non-parametric structural equation model with independent errors (NPSEM, Pearl 2009) and if assumption (1.16) holds, then assumption (1.17) holds if there are no confounders L of the mediator-outcome relationship that are influenced by the treatment themselves. In that case the setting would become more complex, because then these variables L are both confounders and mediators on the causal pathway of interest. Generally, natural direct and indirect effects can not be identified in the presence of so-called *intermediate confounders* (Avin et al. 2005).

The assumption of cross-world independence (1.17) has been a source of much controversy. It implies that mediation analyses are bound to rely on assumptions that can not be guaranteed by study design, not even in randomised cross-over trials of the exposure (Robins and Greenland 1992). Since $Y(a,m)$ and $M(a^*)$ only seem to coexist across multiple worlds, assumption (1.17) can also not be empirically verified, because when a is different from a^* the observed data simply carry no information about the dependence of $Y(a,m)$ and $M(a^*)$.

Under assumptions (1.14) to (1.17), the natural direct and indirect effect are nonparametrically identified as

$$DE(a) = \int\int \{\mathbb{E}(Y_i|M_i = m, A_i = 1, X_i = x) - \mathbb{E}(Y_i|M_i = m, A_i = 0, X_i = x)\} \times F_{M_i|A_i=a,X_i=x}(m)\, F_{X_i}(x)\, dx\, dm \qquad (1.18)$$

and

$$IND(a) = \int\int \mathbb{E}(Y_i|M_i = m, A_i = a, X_i = x) \times \left\{F_{M_i|A_i=1,X_i=x}(m) - F_{M_i|A_i=0,X_i=x}(m)\right\} F_{X_i}(x)\, dx\, dm, \quad (1.19)$$

respectively, using the *mediation formula* (Pearl 2012). This can be seen since assumptions (1.14) to (1.17) and the consistency assumption imply that

$$\begin{aligned}
&\mathbb{E}[Y_i\{a, M_i(a^*)\}] \\
&= \int \mathbb{E}[Y_i\{a, M_i(a^*)\}|X_i = x]\, F_{X_i}(x)\, dx \\
&= \int\int \mathbb{E}[Y_i(a,m)|X_i = x, M_i(a^*) = m]\, F_{M_i(a^*)|X_i=x}(m)\, F_{X_i}(x)\, dx\, dm \\
&= \int\int \mathbb{E}[Y_i(a,m)|X_i = x, M_i(a^*) = m, A_i = a^*]\, F_{M_i(a^*)|X_i=x}(m)\, F_{X_i}(x)\, dx\, dm \\
&= \int\int \mathbb{E}[Y_i(a,m)|X_i = x, A_i = a^*]\, F_{M_i(a^*)|X_i=x}(m)\, F_{X_i}(x)\, dx\, dm \\
&= \int\int \mathbb{E}[Y_i(a,m)|X_i = x, A_i = a]\, F_{M_i(a^*)|X_i=x,A_i=a^*}(m)\, F_{X_i}(x)\, dx\, dm \\
&= \int\int \mathbb{E}[Y_i(a,m)|X_i = x, A_i = a, M_i = m]\, F_{M_i(a^*)|X_i=x,A_i=a^*}(m)\, F_{X_i}(x)\, dx\, dm \\
&= \int\int \mathbb{E}[Y_i|X_i = x, A_i = a, M_i = m]\, F_{M_i(a^*)|X_i=x,A_i=a^*}(m)\, F_{X_i}(x)\, dx\, dm \\
&= \int\int \mathbb{E}[Y_i|X_i = x, A_i = a, M_i = m]\, F_{M_i|X_i=x,A_i=a^*}(m)\, F_{X_i}(x)\, dx\, dm \qquad (1.20)
\end{aligned}$$

where the third equality follows from assumption (1.14), assumption (1.17) and assumption (1.16) are used to establish the fourth and sixth equality respectively, assumption (1.14) together with assumption (1.15) are used to establish the fifth equation and the final equalities follow from the consistency assumption. Changing a and a^* in (1.20) and subtracting the expressions accordingly results in the equations for the natural direct (1.18) and indirect effect (1.19).

Note that under the no-interaction assumption (i.e., $IND = IND(1) = IND(0)$ and $DE = DE(1) = DE(0)$), these natural direct and indirect effects coincide with the Baron and Kenny (1986) approach if the linear regression models of (1.2) and (1.3) are correctly specified (VanderWeele and Vansteelandt 2009; Imai et al. 2010). When linearity assumptions no longer hold, the traditional structural equation models fail (MacKinnon 2008). The advantage of the mediation formula is that it enables one to estimate natural direct and indirect effects via arbitrary models for the mediator and the outcome.

A number of papers have been published on this topic and give closed-form expressions for natural direct and indirect effects for a limited number of mediator and outcome combinations (Lange and Hansen 2011; Valeri and VanderWeele 2013; VanderWeele 2009; VanderWeele and Vansteelandt 2010; VanderWeele 2011). Combining parameter estimates from a regression model for the mediator and a regression model for the outcome in a specific way, however, often results in complex expressions if they can be defined altogether (Lange et al. 2012). In time-to-event scenarios, Lange and Hansen (2011) show how to combine normal regression models for continuous mediators with additive hazard models for event times in a somewhat restrictive setting with no interactions involving the mediator or the exposure. VanderWeele (2011) uses Cox proportional hazards models or accelerated failure time models in the case of rare events. Huang and Cai (2016) achieve greater flexibility by relying on semiparametric probit models for the event-time which combine well with linear models for possibly multiple mediators. Imai et al. (2010) suggest a different, more generic approach based on Monte Carlo sampling to derive natural direct and indirect effects in their R-library *mediation*. It has the advantage that it can handle all types of mediators and outcome variables and that one only has to specify a model for the mediator and the outcome, to get natural direct and indirect effect estimates in terms of mean differences. In case of survival times, reporting effects in terms of mean differences is less ideal, however, because of skewness and censoring. In time-to-event settings, the *mediation* package avoids the need for the rare event assumption that VanderWeele (2011) had to deal with at the cost of demanding parametric survival models. Recently, a lot of work has been done on flexibly implementing these natural direct and indirect effects via natural effect models (Lange et al. 2012; Vansteelandt et al. 2012a; Loeys et al. 2013; Steen et al. 2016). Additionally, there are proposals like the one of Tchetgen Tchetgen (2011) and Zheng and van der Laan (2012a), that offer appealing robustness properties. Considering the concern for bias due to model misspecification, which is quite dominant in the analyses of randomised trials, part of this thesis will focus on robust mediation analysis approaches.

1.2.4 Multiple mediators

Although generally more than one mediational process is of interest or mediators are measured multiple times during the course of the study, a major part of the recent literature on counterfactual-based mediation analysis has been focussed on single mediators measured at one point in time. Simply extending these developments to settings with repeatedly measured or multiple mediators is not possible because of the assumption in (1.17): natural direct and indirect effects can not be identified in the presence of confounders of the mediator-outcome relationship that are influenced by the treatment themselves. In settings with multiple or repeatedly measured mediators, these variables will generally be, next to being mediators, confounders of the mediator - outcome relationship. We saw that the traditional literature on structural equation models (MacKinnon 2008) extended their single mediator approach to multiple or repeatedly measured mediators, but that they work under a strong no-interaction assumption at the individual level, which is biologically rather unlikely and known to be violated in the presence of treatment - mediator interactions and with dichotomous and time-to-event endpoints (Robins and Greenland 1992; Robins 2003; De Stavola et al. 2015). Other work on multiple or repeatedly measured mediators avoids this complication of assumption (1.17) by assuming causally independent multiple mediators (Preacher and Hayes 2008;

Lange et al. 2014; Taguri et al. 2015).

Leaving behind these rather unrealistic assumptions, progress has been made to estimate path-specific effects by jointly modelling the mediators (VanderWeele and Vansteelandt 2013). Imagine a setting with two sequential mediators as in Figure 5. Let $Y(a, M_1(a^*), M_2(a^*, M_1(a^*)))$ denote the counterfactual outcome that would be observed if A were set to a and M_1 and M_2 were set to the value they would have naturally taken if A had been equal to a^*. Then, the total intention-to-treat effect can be decomposed into the joint natural indirect effect (1.21) via the joint mediator (M_1, M_2) from the remaining direct effect of A on Y via neither of the mediators (1.22):

$$
\begin{aligned}
&\mathbb{E}\{Y(1) - Y(0)\} = \\
&\mathbb{E}\{Y(1, M_1(1), M_2(1, M_1(1))) - Y(1, M_1(0), M_2(0, M_1(0)))\} \quad (1.21)\\
&+ \mathbb{E}\{Y(1, M_1(0), M_2(0, M_1(0))) - Y(0, M_1(0), M_2(0, M_1(0)))\} \quad (1.22)
\end{aligned}
$$

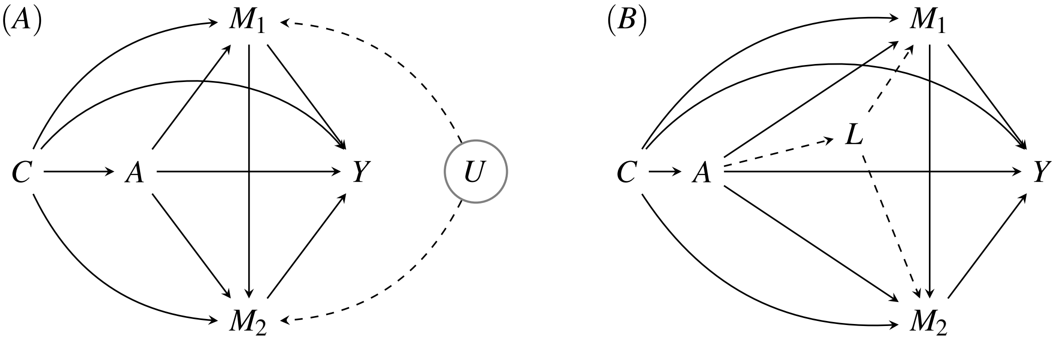


Figure 6: Causal diagram of randomised treatment A, sequential mediators M_1 and M_2, measured confounders C, (unmeasured) confounders of the mediators U and L, and outcome Y.

This two-way decomposition and its joint natural direct and indirect effects are identified if assumptions (1.14) to (1.17) hold for the joint mediator (M_1, M_2). This means that unmeasured common causes U of M_1 and M_2 as in Figure 6a or common causes L that are themselves influenced by the treatment as in Figure 6b are allowed, since U and L are not confounders for the joint effect of (M_1, M_2) on

Y because U and L do not affect Y, except through (M_1, M_2) (VanderWeele and Vansteelandt 2013; Steen et al. 2017).

Going one step further, even a more detailed decomposition can be made. By focusing on M_1 and forgetting M_2 for a moment, the joint natural indirect effect (1.21) can be disentangled into all of the effect going through M_1 (1.23), under the composition assumption that $Y(a, M_1(a^*), M_2(a, M_1(a^*))) = Y(a, M_1(a^*))$, and the remaining effect (1.24) (i.e. only going through M_2):

$$\begin{aligned}
&\mathbb{E}\{Y(1,M_1(1),M_2(1,M_1(1))) - Y(1,M_1(0),M_2(0,M_1(0)))\} = \\
&\quad \mathbb{E}\{Y(1,M_1(1),M_2(1,M_1(1))) - Y(1,M_1(0),M_2(1,M_1(0)))\} \quad (1.23)\\
&\quad + \mathbb{E}\{Y(1,M_1(0),M_2(1,M_1(0))) - Y(1,M_1(0),M_2(0,M_1(0)))\} \, (1.24)
\end{aligned}$$

This means that in Figure 5 two distinct pathways contribute to the natural indirect effect via M_1 ($A \to M_1 \to M_2 \to Y$ and $A \to M_1 \to Y$), while the so-called semi-natural (Pearl, 2014) or partial (Huber, 2014) indirect effect only consists of a single pathway ($A \to M_2 \to Y$). This three-way decomposition is not the only one however. Remember that in the single mediator setting, there were two types (pure and total) of the two path-specific effects (direct and indirect). With an increasing number of mediators, the number of pathways from exposure to outcome and accordingly the number of decompositions of the total intention-to-treat effect grow exponentially. With two mediators there are already eight types of four path-specific effects. This means that more generally in the finest decomposition possible, there are 2^k distinct pathways in a setting with k mediators and $(2^k)!$ possible decompositions (Daniel et al. 2015). The semi-natural or partial indirect effect via M_2 can be identified, similar to the result in (1.21) commonly referred to as Pearl's mediation formula (Pearl 2012), as

$$\begin{aligned}
&\mathbb{E}\{Y(1,M_1(0),M_2(1,M_1(0))) - Y(1,M_1(0),M_2(0,M_1(0)))\} = \\
&\int\int\int \mathbb{E}(Y_i|A_i = 1, M_{1i} = m_1, M_{2i} = m_2, X_i = x)\; F_{M_{1i}|A_i=0,X_i=x}(m_1) \times \\
&\left\{F_{M_{2i}|A_i=1,M_{1i}=m_1,X_i=x}(m_2) - F_{M_{2i}|A_i=0,M_{1i}=m_1,X_i=x}(m_2)\right\} F_{X_i}(x)\; dx\; dm_1\; dm_2,
\end{aligned}$$

Part of the literature (VanderWeele and Vansteelandt 2013; Steen et al. 2017;

Huang and Yang 2017) focusses on decomposing the total intention-to-treat effect into the $k+1$ pathways as described here, since further effect decomposition generally fails. VanderWeele and Vansteelandt (2013) rely on the ordering of the mediators to obtain sequential natural indirect effects. Under the assumptions that the mediators have no unmeasured common causes and that none of the measured common causes of M_1 and M_2 are influenced by the treatment, they obtain 2 out of the $(k+1)!$ decompositions. Similarly, Huang and Yang (2017) introduce closed form expressions for path-specific effects on different effects scales using semiparametric probit, Aalen additive hazard and Cox proportional hazard models. Under slightly stronger identification assumptions, although the difference is very subtle and of little practical relevance in realistic examples, the flexible approach to mediation analysis based on natural effect models of Steen et al. (2017) allows one to recover all $(k+1)!$ decompositions. In the setting where M_2 is the mediator of interest and M_1 is merely a mediator - outcome confounder influenced by the exposure, Miles et al. (2017) showed that the partial (Huber, 2014) indirect effect only via M_2 is still identified when there is unmeasured confounding of the relationship between M_1 and the outcome. This still leaves the assumption of no unmeasured common causes between mediators, however, which will be unlikely to hold in the setting of repeatedly measured mediators. Another limitation of these approaches is their reliance on the causal structure of the mediators (i.e., M_1 influences M_2, but not the other way around) when different mediators are considered at the same time.

VanderWeele, Vansteelandt, and Robins (2014) made progress with a different kind of effect measures, so-called interventional direct and indirect effects. They differ from natural direct and indirect effects because they do not fix the mediator level to be equal to the counterfactual mediator value at level a or a^*, but to a random draw of the distribution of the mediator at exposure level a or a^* given covariate values C. Unless covariates C fully determine the counterfactual values of the mediator, these measures may yield effects of different magnitudes. Natural direct and indirect effects have been criticized because they are defined via cross-world counterfactuals and even randomised controlled studies are not able to give information about them. Additionally, their practical relevance in policy making has been questioned. Interventional direct and indirect effects are not defined in terms of these cross-world counterfactuals and can thus be identified under weaker assumptions (i.e., mediators are allowed to share unmeasured common causes). They are also policy relevant in the sense that they represent the result of fixing or shifting the mediator distribution to the extent that it is affected by the exposure (VanderWeele, 2013). Natural direct and indirect effects are more closely related to the definition of a mechanism however. These interventional effects as proposed in VanderWeele, Vansteelandt, and Robins (2014) also have the disadvantage that they do not sum to the total effect, which hinders interpretation. Vansteelandt and Daniel (2017) overcome this limitation, their proposal for the multiple mediator setting can be used even when the direction of causality is unknown among the different mediators. Different from approaches relying on the causal structure of the mediators (VanderWeele and Vansteelandt 2013; Steen et al. 2017), other pathways that contribute to the indirect effect are identified. The interventional indirect effect via M_1 captures all of the causal pathways from exposure to outcome via M_1 except those involving causal descendants of M_1 (e.g., M_2 in figure 5). The interventional indirect effect via M_1 in Figure 5 is thus represented by the pathway $A \rightarrow M_1 \rightarrow Y$, but not $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ unlike the semi-natural indirect effect (Pearl, 2014).

$$\begin{aligned}
&\mathbb{E}\{Y(1,M_1(1),M_2(1,M_1(1))) - Y(1,M_1(0),M_2(1,M_1(0)))\} = \\
&\qquad \mathbb{E}\{Y(1,M_1(1),M_2(1,M_1(1))) - Y(1,M_1(1),M_2(1,M_1(0)))\} \quad (1.25) \\
&\qquad + \mathbb{E}\{Y(1,M_1(1),M_2(1,M_1(0))) - Y(1,M_1(0),M_2(1,M_1(0)))\} \quad (1.26)
\end{aligned}$$

Further disentangling the natural indirect effect only via M_1 $(A \rightarrow M_1 \rightarrow Y)$ from the pathway $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ is impossible. These effects not identifiable (Avin et al. 2005; Daniel et al. 2015) unless one is willing to make certain no-interaction assumptions (Huber, 2014; Imai and Yamamoto, 2013; Petersen et al., 2006; Robins, 2003; Tchetgen Tchetgen and VanderWeele, 2014) or makes use of sensitivity analysis (Albert and Nelson 2011; Daniel et al. 2015; Imai and Yamamoto 2013). Looking at expressions (1.25) and (1.26) shows that further disentangling the effect (1.23) makes little sense because this would require knowledge about counterfactual outcome $Y(1, M_1(1), M_2(1, M_1(0)))$ which makes use of the counterfactual mediator M_1 in both intervention worlds. Figures 5 and 6 show that we can not just forget about M_1 and focus on M_2 alone because M_1 is a confounder of the relationship between M_2 and the outcome Y.

The previously discussed literature, all handles settings with a single exposure and outcome, together with two or more mediators or a single repeatedly measured mediator. It does not allow the exposures and the multiple mediators to vary over time. Similar to VanderWeele et al. (2014) and Vansteelandt and Daniel (2017), VanderWeele and Tchetgen Tchetgen (2017) and Lin et al. (2016) make progress via interventional effects. The disadvantage of their interventional direct and indirect effects in the presence of time-varying exposures, mediators and confounders is that they do not always sum to the total intention-to-treat effect, which makes interpretation difficult. An additional limitation of their proposal is the way they account for confounding in their mediational g-formula. They intend to estimate, for example, the interventional analogue of the natural indirect effect as the change in outcome under randomised exposure a if the mediator for each individual were fixed to a random draw from the distribution of the mediator under randomised exposure a^*. This random draw from the mediator distribution under exposure a^* only depends on patients' observed baseline confounders however and thus ignores the observed time-varying confounders. As a result, the values for the mediator might really deviate from the mediator trajectory a person would have “naturally” followed over time. Zheng and van der Laan (2017) overcome this limitation within the interventional direct and indirect effects setting for time-varying exposures and allow patient's time-varying covariate data to influence the random draw from the mediator distribution. As a result of taking these time-varying confounders into account, their interventional indirect effect captures the effect of treatment on outcome as transmitted along the combination of pathways whereby treatment directly influences one of the mediators, and not those pathways whereby treatment initially influences one of the time-varying confounders.

Moet dit gedetailleerder? Johan is precies de enige die zo in detail gaat?

1.3 Outline and contributions

In the detailed and up-to-date overview of the causal inference literature on mediation analysis in this chapter, we saw that analyses of randomised experiments can often benefit from a mediation analysis next to the primary intention-to-treat analysis. As the literature of applied papers on the topic in medical and epidemiological journals grows, more and more applied researchers recognize the usefulness of mediation analyses. As such, an increasing number of analyses of randomised trials include an attempt to further examine the treatment mechanism and try to decompose the intention-to-treat effect into an indirect effect, mediated by given intermediaries and the remaining direct effect. Although the traditional approach to direct and indirect effects may be straightforward and intuitive, extensions to non-linear models are limited for a number of earlier discussed reasons. Novel developments from causal mediation analysis led to the formal decomposition of direct and indirect effects even in settings with interactions and nonlinearities. The main identification result of this formal mediation analysis framework, referred to as the *mediation formula* (Pearl 2001, 2012), allows the combination of arbitrary models for the outcome and the mediator. Mediation analyses based on a direct application of the mediation formula have a major disadvantage however: the results may not be unbiased if one of the models, the one for the outcome or the mediator, is misspecified. Considering the concern for bias due to model misspecification, which is dominant in analyses of randomised controlled trials, more robust approaches approaches that are less sensitive to model misspecification are considered in **chapter 2** and **chapter 3**.

One alternative strategy to the mediation formula (Tchetgen Tchetgen and Shpitser 2012; Vansteelandt 2012b) is substituting the model for the mediator or the model for the outcome with a model for the exposure. This is typically done when the exposure is randomly assigned as one can then be sure about the correctness of the exposure model. In **chapter 2**, we focus on such robust mediation analysis approaches for continuous and binary outcomes, in the sense that we exploit a priori knowledge of the randomisation probabilities and only require correct specification of the model for the outcome. The model for the outcome, and not that of the mediator, because it avoids the need for inverse probability weighting and because a model for the mean outcome is generally more easily specified than a model for the mediator distribution (as needed in the mediation formula). Unfortunately, not relying on a model for the distribution of the mediator can result in a considerable loss of efficiency when baseline covariates that are predictive of the mediator are available. As such, we propose a novel approach that makes use of a model for the mediator to extract all available information from baseline covariates, but still delivers unbiased results if this model is misspecified. To supplement this chapter (based on the paper of Vandenberghe et al. (2017a)), we developed an R-function which is available as part of the supplementary material of the paper and allows researchers to make use of the proposed estimators. In **chapter 3**, we describe similar modern mediaton analysis techniques for time-to-event endpoints and extend the estimators proposed in **chapter 2** to a survival setting. Although the proposed analysis strategy is broadly applicable to mediation analyses of time-to-event endpoints, in this chapter we focus on showing how the analysis may be informative with respect to pathological complete response (pCR) as a putative surrogate marker when data from just a single trial are available. Again, we supplemented this chapter (based on the paper of Vandenberghe et al. (In press)) with an easy-to-use R-function which is also available as part of the supplementary material of the paper.

Current mediation analysis approaches are generally focussed on the effect of treatment on outcome via a single mediator measured at a single point in time. This is rather limiting since more mediators are often of interest and/or measured multiple times during the course of a study. In **chapter 4**, we give an overview of the existing advances on the topic of longitudinal mediation analysis and propose a mediation analysis strategy for a randomly assigned exposure, a repeatedly measured mediator and a time-to-event endpoint in the presence of (repeatedly measured) time-varying confounders. We make progress, similar to recent contributions of Zheng and van der Laan (2012b) and Zheng and van der Laan (2017), but focus on natural direct and indirect effects (in contrast to interventional direct and indirect effects) for a time-to-event outcome. We conclude in **chapter 5** with a discussion on the relevance of this research in the fast-growing field of mediation analysis and discuss some further challenges.

CHAPTER 2

Efficient mediation analyses of binary and continuous outcomes

This chapter is based on the following paper: Vandenberghe, S., Vansteelandt, S., and Loeys, T. (2017). Boosting the precision of mediation analyses of randomised experiments through covariate adjustment. *Statistics in Medicine, 36(6),* 939-957.

Analyses of randomised experiments frequently include attempts to decompose the intention-to-treat effect into a direct and indirect effect, mediated by given intermediaries, with the aim to shed light onto the treatment mechanism. Methods from causal mediation analysis have facilitated this by allowing for arbitrary models for the outcome and the mediator. They thereby generalise the traditional approach to direct and indirect effects, which is essentially limited to linear models. The default maximum likelihood methods make use of a model for the conditional distribution of the mediator, given treatment and baseline covariates, but are prone to bias when that model is misspecified. In randomised experiments, specification of such model can be easily avoided, but at the expense of a sometimes major efficiency loss when those baseline covariates are predictive of the mediator. In this article, we develop a compromise approach: it makes use of a model for

the mediator to optimally extract information from the baseline covariate data, but is insulated from the impact of misspecification of that model; it achieves this by exploiting the known randomization probabilities. Simulation studies and the analysis of a randomised study show major efficiency gains and confirm our theoretical findings that the default methods from causal mediation analysis are sometimes, though not always, reasonably robust to model misspecification.

2.1 Introduction

Analyses of randomised experiments frequently supplement the primary intention-to-treat analysis with analyses aimed at a better understanding of the treatment mechanism. Mediation analyses seek a more in-depth understanding by decomposing the intention-to-treat effect into a direct and indirect effect, mediated by given intermediaries (Emsley et al. 2010). For instance, Oba et al. (2011) contrasted two treatments that equally suppressed the incidence of cardiovascular events, but had a small but significantly different effect on systolic blood pressure. A mediation analysis provided a more in-depth understanding by clarifying what the (relative) treatment effect would be if an effect on systolic blood pressure could be avoided. Rosenblum et al. (2009) observed a much lower reported use of condoms in the intervention arm than in the control arm of the open-label ‘Methods for improving reproductive health in Africa’ trial that investigated the effect of diaphragm and lubricant gel use in reducing HIV infection among susceptible women. They used mediation analysis to assess what the effect of diaphragm and lubricant gel use would have been, had a harmful effect on condom use been avoided.

Methods from causal mediation analysis have generalised traditional mediation analysis approaches, which are inspired by linear structural equation models (Baron and Kenny 1986). The key identification result underlying these methods - the so-called mediation formula (Pearl 2001; VanderWeele and Vansteelandt 2009) - suggests possibilities to combine arbitrary models for the outcome and mediator. In particular, maximum likelihood estimates (MLE) of direct and indirect effects are obtained by averaging predicted values from a model for the conditional outcome mean, given mediator and covariates and given a specific treatment level, over the conditional distribution of the mediator at a given (possibly different) treatment level, given those covariates; such averaging is much akin to the use of standardisation in epidemiology. A major concern with such mediation analyses is that they may deliver biased results when those models for the outcome and mediator are misspecified.

This concern for model misspecification is very pertinent in intention-to-treat analyses that rely on statistical models to adjust for baseline covariates (Pocock et al. 2002; Rosenblum and van der Laan 2009). Since model-free analyses are usually attainable here, covariate-adjusted intention-to-treat analyses of the overall treatment effect are often viewed with some scepticism, despite the precision benefits that they may confer (Pocock et al. 2002; Senn 2000). Recent developments have nevertheless given rise to an array of methods that make use of statistical models to boost the precision of the intention-to-treat analysis, but are not susceptible to bias when those models are misspecified (Tsiatis et al. 2008; Zhang et al. 2008; Moore and van der Laan 2009; Colantuoni and Rosenblum 2015; Vermeulen et al. 2015). The precision benefit of these analyses is derived from the potential of covariate adjustment to eliminate noise from the outcome data; their robustness against model misspecification is secured by exploiting the a priori knowledge of the randomisation probabilities.

Model-free mediation analyses are not feasible, however, because of the need to control for confounding of the mediator-outcome association. In the usual settings where the confounders are continuous and/or discrete with many levels, this demands specification of either a model for the outcome or a model for the mediator (Tchetgen Tchetgen and Shpitser 2012; Vansteelandt 2012b). In this article we will focus on strategies that do not rely on models for the mediator for unbiased estimation, for two reasons. First, models for the *distribution* of the mediator are generally more difficult to specify than models for the *mean* of the outcome. Second, our focus on outcome models will avoid the need for inverse probability weighting, which can sometimes yield estimators with erratic behaviour that may moreover be sensitive to minor misspecifications in the tails of the mediator distribution

(Vansteelandt 2012b). In particular, the semi-parametric efficient strategies that we will propose, make use of a mediator model, but only to extract information from the baseline covariate data; the proposed estimators will be insulated from the impact of misspecification of that mediator model. By construction, they thus form an ideal compromise between maximum likelihood-based mediation analyses that make use of a mediator model, and simple alternatives that do not. The performance of the proposed estimators, relative to various alternatives, is investigated through simulation studies and through the analysis of a randomised experiment on the effect of implicit priming.

2.2 Assessing mediation in randomised experiments

2.2.1 Definitions and assumptions

Consider a study design which collects baseline covariates X_i (e.g. age, gender, ...) for a random sample of individuals ($i = 1,...,n$), who are subsequently randomised over experimental treatment ($A_i = 1$) or control treatment ($A_i = 0$), with known randomisation probabilities that possibly depend on covariates. For each individual, data are recorded on a potential mediator M_i - which may consist of multiple components - and the end-of-study outcome Y_i. The direct effect of treatment on outcome quantifies that part of the treatment effect which is not mediated by M_i. To be precise about its meaning, we make use of potential outcome notation. In particular, let M_{i0} and Y_{i0} denote the value that the mediator and outcome would have taken for individual i, had this individual been assigned to the control arm; M_{i0} and Y_{i0} equal the observed value of the mediator and outcome for control individuals, but remain unobserved for individuals on the treatment arm. Further, let $Y_{i1M_{i0}}$ be the value that the outcome would have taken for individual i, had this individual been assigned to the treatment arm, but his level of the mediator were as it would have been under control conditions; $Y_{i1M_{i0}}$ is unobservable for all individuals, but enables us to formally express the direct effect of treatment on outcome as

$$E\left(Y_{i1M_{i0}} - Y_{i0}\right). \tag{2.1}$$

Rewriting Y_{i0} as $Y_{i0M_{i0}}$, it is indeed seen that the above measure captures the effect of treatment, while holding the mediator fixed at a level M_{i0}. This level M_{i0} is often regarded as a natural mediator level for the given individual in the sense that it represents the level that would be seen on the control arm; expression (2.1) is therefore referred to as a natural direct effect (Robins and Greenland 1992; Pearl 2001; VanderWeele and Vansteelandt 2009). A natural indirect effect is likewise obtained as

$$E\left(Y_{i1} - Y_{i1M_{i0}}\right). \tag{2.2}$$

Rewriting Y_{i1} as $Y_{i1M_{i1}}$, it is indeed seen that the above measure captures the indirect effect of treatment because it expresses the effect of changing the level of the mediator only to the extent that it is affected by treatment: that is, changing it from M_{i0} to M_{i1}. Note that these natural direct and indirect effects sum to the total treatment effect $E\left(Y_{i1} - Y_{i0}\right)$.

Randomisation in itself does not suffice to disentangle direct from indirect treatment effects (Pearl 2001; VanderWeele and Vansteelandt 2009). The reason is that the magnitude of the indirect (and hence direct) effect depends on how strongly the mediator affects the outcome; since the mediator is not randomly assigned, estimating this requires knowledge of all confounders of the relation between mediator and outcome. Throughout, we will therefore assume that the baseline covariate set X_i is sufficient to adjust for confounding of this association. In some studies, some of the confounders of the relation between mediator and outcome will only manifest themselves as the study is ongoing, and may then be themselves influenced by the treatment. The results that we propose in this article cannot handle such so-called intermediate confounders; we refer the reader to Vansteelandt and VanderWeele (2013), VanderWeele et al. (2014) or Tchetgen Tchetgen and VanderWeele (2014) for strategies to deal with intermediate confounders and to the Appendix for formal identification assumptions.

2.2.2 Traditional approaches for linear models

When the outcome obeys a linear regression model with only additive effects, e.g.

$$E(Y|A,M,X) = \beta_0^* + \beta_1^* A + \beta_2^* M + \beta_3^{*t} X,$$

then the natural direct effect (2.1) can be calculated as β_1^* (Baron and Kenny 1986; VanderWeele and Vansteelandt 2009). The natural indirect effect can then be calculated in two standard ways. The first is to fit the regression model

$$E(M|A) = \alpha_0^* + \alpha_1^* A, \tag{2.3}$$

and calculate the indirect effect as $\alpha_1^*\beta_2^*$ (Baron and Kenny 1986; VanderWeele and Vansteelandt 2009). The second, more standard strategy is to fit the regression model

$$E(M|A,X) = \gamma_0^* + \gamma_1^* A + \gamma_2^{*t} X, \tag{2.4}$$

and calculate the indirect effect as $\gamma_1^*\beta_2^*$. Estimates for these direct and indirect effects can now be obtained by substituting $\alpha_1^*, \beta_1^*, \gamma_1^*$ and β_2^* by ordinary least squares estimates. Since model (2.3) is always correctly specified (by virtue of A being dichotomous), the resulting direct and indirect effects are unbiased (in large samples) as soon as the linear outcome model is correctly specified. When the exposure is randomly assigned independently of X, then interestingly, this is also the case for the direct and indirect effects obtained under model (2.4), in spite of the fact that this model may be misspecified. That these direct and indirect effect estimates are insulated from the impact of misspecification of the mediator model, follows from the properties of ordinary least squares estimators and the fact that A and X are orthogonal predictors (in the sense that they are independent, by design) (Yang and Tsiatis 2001; Rosenblum and van der Laan 2009). It moreover follows from the properties of ordinary least squares estimation that the resulting direct and indirect effect estimators are not less efficient than those obtained under model (2.3), even when model (2.4) is misspecified (Yang and Tsiatis 2001). In view of this, we conclude that the direct and indirect effect estimators obtained under model (2.4) are preferred. However, they will not be fully efficient when model (2.4) is misspecified. The results in the next sections accommodate this, and will moreover generalise this result to non-linear outcome and mediator models.

2.2.3 Causal mediation analysis approaches for linear and non-linear models

The causal mediation literature has extended the traditional approaches from the previous section to arbitrary models for outcome and mediator. We will here review two relatively standard causal mediation analysis approaches to estimate $E\left(Y_{i1M_{i0}}\right)$ and $E\left(Y_{i0}\right)$, and thus the natural direct effect as their difference. The estimation of $E\left(Y_{i0M_{i1}}\right)$ and $E\left(Y_{i1}\right)$, needed for the calculation of the natural indirect effect, then follows by simply recoding the exposure.

The estimation of $E\left(Y_{i1M_{i0}}\right)$ generally demands modelling assumptions (Tchetgen Tchetgen and Shpitser 2012; Vansteelandt 2012b). As before, we will assume throughout that a correctly specified model is available for the outcome in the treatment arm, in function of mediator M and confounders X. This may be a linear regression model, e.g.

$$E(Y|A=1,M,X) = \eta_0^* + \eta_1^* M + \eta_2^{*t} X, \tag{2.5}$$

or a logistic regression model, e.g.

$$E(Y|A=1,M,X) = \text{expit}\left(\eta_0^* + \eta_1^* M + \eta_2^{*t} X + \eta_3^{*t} MX\right). \tag{2.6}$$

More generally, we shall assume that $E(Y|A=1,M,X)$ obeys the model

$$E(Y|A=1,M,X) = m(M,X;\eta^*), \tag{2.7}$$

where $m(M,X;\eta)$ is a known function, evaluated at an unknown parameter η^*; e.g. $m(M,X;\eta) = \eta_0 + \eta_1 M + \eta_2^t X$ in model (2.5) and $m(M,X;\eta) =$ $\text{expit}\left(\eta_0 + \eta_1 M + \eta_2^t X + \eta_3^t MX\right)$ in model (2.6). Let $\hat{\eta}$ be an estimator of η^*, as obtained using a standard regression procedure.

The most popular approach for estimating $E\left(Y_{i1M_{i0}}\right)$ is based on maximum likelihood estimation (MLE). It involves taking the average of the fitted values $m(M_i,X_i;\hat{\eta})$ from the outcome model (with $\hat{\eta}$ the MLE) over the fitted distribution of M_i in the untreated ($A_i = 0$), given X_i (based on MLE under some parametric model) and then averaging the result across all individuals as follows (VanderWeele and Vansteelandt 2009; Imai et al. 2010):

1. Fit a parametric model for the distribution of M in the untreated ($A = 0$), given X, using MLE. For each subject i, both treated and untreated, take a large number K (e.g. $K = 10000$) of random draws $M_{i(1)}, ...M_{i(K)}$ from this distribution with X set to X_i.

2. Estimate $E\left(Y_{i1M_{i0}}\right)$ as

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{K}\sum_{j=1}^{K} m(M_{i(j)}, X_i; \hat{\eta})\right\}.$$

Note that the averaging is across all subjects, which is valid by the fact that treated and untreated subjects are exchangeable by randomisation. The outcome mean $E\left(Y_{i0}\right)$ under control treatment can be estimated similarly upon redefining $m(M, X; \eta)$ to be a parametric model for the expected outcome $E(Y|A = 0, M, X)$ in the untreated. This approach, which generalises the strategy based on model (2.4) in the previous section, has been implemented in a number of software packages, e.g. 'mediation' in R (Imai et al. 2010). A drawback is that it requires specification of a model for the mediator. This raises concern for bias in the resulting estimates for $E\left(Y_{i1M_{i0}}\right)$ when that model is misspecified.

With concern for bias, a preferable approach for the analysis of randomised experiments may therefore be to avoid reliance on a model for the mediator. The outcome mean $E\left(Y_{i0}\right)$ under control treatment can be estimated as the average outcome in the control arm, which is valid by randomisation. Furthermore, $E\left(Y_{i1M_{i0}}\right)$ can be estimated as the average of the fitted values $m(M_i, X_i; \hat{\eta})$ across individuals in the control arm (Tchetgen Tchetgen and Shpitser 2012; Albert 2012; Vansteelandt 2012b):

$$\frac{1}{n_0}\sum_{i:A_i=0} m(M_i, X_i; \hat{\eta}); \tag{2.8}$$

with n_0 the number of control subjects; e.g., this is

$$\frac{1}{n_0}\sum_{i:A_i=0}(\hat{\eta}_0+\hat{\eta}_1 M_i+\hat{\eta}_2 X_i)$$

under model (2.5), and

$$\frac{1}{n_0}\sum_{i:A_i=0}\text{expit}\left(\hat{\eta}_0+\hat{\eta}_1 M_i+\hat{\eta}_2^t X_i+\hat{\eta}_3^t M_i X_i\right)$$

under model (2.6). That this works can be intuitively understood upon noting that $m(M,X;\,\eta^*)$ represents the average outcome in the treated at their observed mediator and covariate values; thus by averaging across control subjects, this gets evaluated at control levels, M_{i0}, for the mediator. Limiting the average to the controls is justified since both groups are comparable by randomisation. This estimator of $E\left(Y_{i1M_{i0}}\right)$, which we will refer to as the restricted MLE (RMLE), generalises the strategy based on model (2.3) in the previous section. Like the considered estimator of $E\left(Y_{i0}\right)$ it is easy to obtain, but does not exploit all information in the data, by primarily or exclusively relying only on data for control subjects and thus ignoring that the treated subjects represent the same population. In the next section, we will improve the efficiency of these estimators by additionally making use of information obtained for treated subjects. In particular, we will infer the semi-parametric efficient estimator of $E\left(Y_{i1M_{i0}}\right)$ under model (2.7) and of $E\left(Y_{i0}\right)$ under the nonparametric model.

2.2.4 Proposal

Efficient estimation of the natural direct and indirect effects requires the specification of three models in addition to the outcome model in the treated. However, their specification will turn out to be relatively innocent because one model will turn out to be a priori known by design, and misspecification of the two others will turn out not to induce bias (although it may affect precision).

The first model is a so-called propensity score model for the probability of treatment, given covariates. Specifying such model is straightforward in randomised experiments where the randomisation probabilities are known by design. For instance, we may assume that

$$P(A=1|X)=\alpha^*,$$

or that

$$P(A=1|X)=\text{expit}\left(\alpha_0^*+\alpha_1^{*t}X\right).$$

Both these models are correctly specified models when, as often, all individuals are equally likely to be administered treatment or control, regardless of their covariates (i.e. $P(A=1|X)=0.5$). More generally, and for notational convenience, we shall assume that $P(A=1|X)$ obeys the model

$$P(A=1|X)=g(X;\alpha^*),$$

where $g(X;\alpha)$ is a known function, evaluated at an unknown parameter α^* (e.g., $g(X;\alpha)=\alpha$ or $g(X;\alpha)=\text{expit}\left(\alpha_0+\alpha_1^tX\right)$ in the above examples). Here, α^* can be estimated (as $\hat{\alpha}$) using a default maximum likelihood procedure. By allowing for a possible dependence on covariates, our proposed strategy, unlike that in Section 2.2.2, will be able to accommodate designs where treatment is randomised with randomisation probabilities depending on measured covariates.

The second model, a regression model for the outcome on covariates in the control arm, is only needed for efficient estimation of $E\left(Y_{i0}\right)$, as in Tsiatis et al. (2008). This may be a linear regression model, e.g.

$$E(Y|A=0,X)=\beta_0^*+\beta_1^{*t}X$$

or a logistic regression model, e.g.

$$E(Y|A=0,X)=\text{expit}\left(\beta_0^*+\beta_1^{*t}X\right).$$

More generally, we will consider generalised linear models of the form

$$E(Y|A=0,X)=z(\beta_0^*+\beta_1^{*t}X), \tag{2.9}$$

where $z(.)$ is a canonical link function (e.g. the identity link, the exponential link, the inverse logistic transformation, ...). Here, X may be a different vector of baseline covariates than considered in model (2.7) (e.g. it may include outcome predictors that do not confound the mediator-outcome association and were therefore not included in model (2.7)). We will then fit the model using the default maximum likelihood procedure for generalised linear models on data for the controls, weighing each observation by 1 over the fitted probability of being assigned to the control arm, given covariates; such weighing is not required when the randomisation probabilities are constant (i.e., $g(X;\alpha) = \alpha$). Let $\hat{\beta}$ be the resulting estimator of $\beta^* \equiv (\beta_0^*, \beta_1^{*t})^t$. Choosing a canonical link function, limits the class of models, but can be justified in that it will lead to a simple estimation procedure and, moreover, misspecification of this model will turn out not to induce bias. In the Appendix, we show how to conduct inference under a larger class of models.

The third model, which is only needed for efficient estimation of $E\left(Y_{i1M_{i0}}\right)$, regresses the fitted values from the outcome model for $E(Y|A=1,M,X)$ on covariates in the control arm. This model thereby produces predictions of $Y_{i1M_{i0}}$ that depend solely on the available baseline covariates, and which can therefore be evaluated for *all* individuals (not just control subjects, as was the case for the RMLE). For instance, reconsidering models (2.5) and (2.6) for scalar M and X, we may assume that

$$E\left(\eta_0^* + \eta_1^* M + \eta_2^{*t} X | A=0, X\right) = \gamma_0^* + \gamma_1^{*t} X,$$

which is satisfied if the mean of the mediator is linear in covariates X within the controls, or that

$$E\left\{\text{expit}\left(\eta_0^* + \eta_1^* M + \eta_2^{*t} X + \eta_3^{*t} MX\right) | A=0, X\right\} = \text{expit}\left(\gamma_0^* + \gamma_1^{*t} X\right).$$

More generally, we will consider generalised linear models of the form

$$E\left\{m(M,X;\eta^*)|A=0,X\right\} = z(\gamma_0^* + \gamma_1^{*t} X), \tag{2.10}$$

where $z(.)$ is again a canonical link function and where, again, X may be a different vector of baseline covariates than considered in models (2.7) and (2.9). An estimator $\hat{\gamma}$ of $\gamma^* \equiv (\gamma_0^*, \gamma_1^{*t})^t$ is then obtained as in the previous paragraph, via maximum likelihood on data in the control arm, weighing each observation by 1 over the fitted probability of being assigned to the control arm, given covariates. In the Appendix, we show how to do inference under a larger class of models.

Once these models are fitted, an estimator of $E\left(Y_{i0}\right)$ and $E\left(Y_{i1M_{i0}}\right)$ is obtained by simple averaging of the fitted values of model (2.9) and (2.10) across all individuals, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} z(\hat{\beta}_0 + \hat{\beta}_1^t X_i), \tag{2.11}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} z(\hat{\gamma}_0 + \hat{\gamma}_1^t X_i), \tag{2.12}$$

respectively. These estimators improve upon the efficiency of the average outcome in the control arm, and of the restricted MLE from Section 2.2.3, respectively. This can be intuitively understood in that they are obtained by averaging predictions for all subjects, and not just for the controls (as in (3.5), for instance). More precisely, it follows from Tsiatis et al. (2008) that the estimator (2.11) is efficient when the second working model is correctly specified, but remains unbiased for $E\left(Y_{i0}\right)$ in large samples, even when that model is misspecified. It follows from the Appendix that the estimator (2.12) is efficient when the third working model is correctly specified, but remains unbiased for $E\left(Y_{i1M_{i0}}\right)$ in large samples, even when that model is misspecified. Note, however, that the estimator (2.12) does require the outcome model for $E(Y|A=1,M,X)$ to be correctly specified. In the Appendix, we show how standard errors can be calculated and further show that, even while α^* is generally known in randomised experiments, estimating it using maximum likelihood improves efficiency.

The estimator (2.12) is very closely related to the MLE of $E\left(Y_{i1M_{i0}}\right)$, which is obtained by averaging the fitted outcome means $m(M_i, X_i; \hat{\eta})$ over the fitted distribution of the mediator in the controls, given covariates, under some parametric mediator model. In particular, when the MLE $\hat{\delta}$ of the parameters δ that index the parametric mediator model satisfies

$$\int m(u,X;\hat{\eta})f(M=u|A=0,X;\hat{\delta})du = z(\hat{\gamma}_0+\hat{\gamma}_1^{*t}X),$$

then the MLE of $E\left(Y_{i1M_{i0}}\right)$ is mathematically identical to (2.12), provided that the simple exposure model $P(A=1|X)=\alpha^*$ is used (so that the weights equalling 1 over the fitted probability of being assigned to the control arm, given covariates, are constant). This is obviously the case when $m(M,X;\eta^*)$ does not depend on M (in which case no mediator model is needed), as well as when the linear model (2.5) is combined with a normal regression model for the mediator in the untreated, with mean also linear in X. It motivates why, as we will see in simulation studies, the MLE is sometimes reasonably robust to model misspecification.

2.2.5 Improvements that guarantee efficiency gain, even under model misspecification

A drawback of the estimators of $E\left(Y_{i0}\right)$ and $E\left(Y_{i1M_{i0}}\right)$ that we proposed in Section 2.2.4 is that their efficiency is only guaranteed under correct specification of model (2.9) and (2.10), respectively; under misspecification of these models, the proposed estimators may sometimes be less efficient than the simple estimators from Section 2.2.3. We will therefore label these estimators as being locally efficient (LE) in that their efficiency is local: attained under a correctly specified model for (2.9) or (2.10), but not necessarily otherwise.

Easy-to-compute estimators of $E\left(Y_{i0}\right)$ and $E\left(Y_{i1M_{i0}}\right)$ can however be obtained that are at least as efficient as the corresponding restricted MLEs, regardless of correct specification of working models (2.9) and (2.10). For $E\left(Y_{i0}\right)$, this has been noted in Moore and van der Laan (2009) and Yang and Tsiatis (2001) in the context of linear or logistic models when $\alpha^*=0.5$; we here extend it to more general models and to the estimation of $E\left(Y_{i1M_{i0}}\right)$. In particular, when the randomisation probabilities are assumed constant (in the sense that $g(X;\alpha^*)=\alpha^*$) and the working models (2.9) and (2.10) are linear and include an intercept and covariates X, then this is possible via the estimation procedure of Section 2.2.4, but with $\hat{\beta}$ and $\hat{\gamma}$

substituted by estimators $\tilde{\beta}$ and $\tilde{\gamma}$, where:

- $\tilde{\beta}$ is the ordinary least squares estimator obtained upon linearly regressing $\tilde{Y} = (1-A)Y + \hat{E}(Y|A=0)\{A-\hat{\alpha}\}$ onto $\tilde{X} = \{X - \hat{E}(X)\}\{A-\hat{\alpha}\}$ (without intercept), where $\hat{E}(Y|A=0)$ is the sample average of Y in the control group and $\hat{E}(X)$ is the sample average of X;

- $\tilde{\gamma}$ is the ordinary least squares estimator obtained upon linearly regressing $\tilde{M} = (1-A)m(M,X;\hat{\eta}) + \hat{E}\{m(M,X;\hat{\eta})|A=0\}\{A-\hat{\alpha}\}$ onto $\tilde{X}$, where $\hat{E}\{m(M,X;\hat{\eta})|A=0\}$ is the sample average of $m(M,X;\hat{\eta})$ in the control group.

That this procedure delivers estimators that are at least as efficient as the corresponding restricted MLEs is formally shown in the Appendix; it essentially follows from the properties of ordinary least squares estimation, which guarantees residuals with reduced variance, regardless of correct model specification. The resulting estimators may however be less precise than the corresponding estimators of Section 2.2.4 when models (2.9) and (2.10), respectively, are in fact non-linear and correctly specified. We will therefore label these estimators as being restricted efficient (RE) in that their efficiency is restricted to linear models, but global (i.e. not local). More general results for non-linear working models are obtainable using empirical efficiency maximisation (Rubin and van der Laan 2008; Cao et al. 2009), but fall beyond the scope of this chapter.

2.2.6 Improvements that yield more robustness to model misspecification

Tchetgen Tchetgen and Shpitser (2012) and Lendle et al. (2013) also propose natural direct and indirect effect estimators that are applicable to randomised experiments. Their estimators are unbiased for $E(Y_{i1M_{i0}})$ (in large samples) when either the outcome model (2.7) is correctly specified, or a model for the distribution of the mediator, given treatment and covariates (or instead of this, a model for the probability of treatment, given mediator and covariates). They thus have even greater robustness to model misspecification than the estimators of the previous section, but are generally less efficient by not explicitly relying on correct specification of the outcome model (2.7) (even though they are (locally) efficient in the less restrictive nonparametric model). Additional robustness is built in the proposal of Lendle et al. (2013) by making using data-adaptive statistical learning methods. Our choice to develop a different estimation approach in Section 2.2.4 was guided by (a) its much greater simplicity, by (b) the fact that - unlike the estimators in Tchetgen Tchetgen and Shpitser (2012) and Lendle et al. (2013) - it does not require inverse probability weighting when the randomisation probabilities are constant, as weighting can sometimes yield erratic behaviour, and by (c) the fact that we believe models for the *distribution* of the mediator (or the probability of treatment, given mediator and covariates) are generally more difficult to specify than models for the *mean* of the outcome. We refer the reader to Lendle et al. (2013) for details on these Targeted Maximum Likelihood Estimators (TMLEs) of the natural direct effect, and to the Appendix for corresponding TMLEs of the indirect effect. We will evaluate these estimators in the simulation studies of the next section.

2.3 Simulation study

We evaluate the performance of the different proposed estimators through simulation analyses with 1000 runs for data sets of 500 observations. Settings are shown for both continuous and binary outcomes, and a continuous mediator. Additional simulation results for a binary mediator, small sample sizes (i.e. $n = 50$) and misspecified outcome models are reported in the Appendix. In each simulation study, a dichotomous exposure A is drawn with $P(A = 0) = P(A = 1) = 0.5$. We evaluate eight estimators: the restricted MLE (RMLE) and the MLE estimator of Section 2.2.3, with and without treatment interactions, the locally efficient (LE) estimator of Section 2.2.4, the restricted efficient (RE) estimator of Section 2.2.5 and three targeted maximum likelihood estimators (TMLE). The ‘parametric’ TMLE uses a correct model for the outcome and a parametric model for the treatment with main effects of the mediator en confounders. The second ‘partially parametric’ TMLE uses the correct parametric model for the outcome and a library of data-adaptive algorithms for the treatment model, while the third ‘non-parametric’ TMLE uses this library of data-adaptive algorithms (GLM, Step, GLM interaction and Step interaction) for fitting both models (Lendle et al. 2013). For the MLE and the 'parametric' TMLE, we report bootstrap standard errors, while sandwich standard errors (see the Appendix) are presented for all other estimators and used to construct 95% confidence intervals.

2.3.1 Correct model specification

2.3.1.1 Continuous Outcome

Covariates $X = (X_1,...,X_8)^t$ are generated as follows: $X_1, X_3, X_8 \sim \mathcal{N}(0,1)$; X_4 and X_6 are Bernoulli with $P(X_4 = 1) = 0.3$ and $P(X_6 = 1) = 0.5$; and $X_2 = 0.2X_1 + 0.98U_1$, $X_5 = 0.1X_1 + 0.2X_3 + 0.97U_2$, and $X_7 = 0.1X_3 + 0.99U_3$, where $U_l \sim \mathcal{N}(0,1), l = 1,2,3$. The continuous mediator M is drawn from a normal distribution with residual variance 1 and mean $E(M|A,X) = \beta_0 + \beta_1 A + \beta_2^t X$ with $\beta_0 = 0$, $\beta_1 = 1$ and β_2 chosen to yield null, moderate, and strong associations between mediator M and covariates X, as detailed in the Web Appendix. Next, a continuous outcome Y is drawn from a normal distribution with mean $E(Y|A,M,X) = \alpha_0 + \alpha_1 A + \alpha_2 M + \alpha_3^t X$ and residual variance 4. Here, $\alpha_0 = 1$, $\alpha_1 = 2$, α_2 is set to -1 and α_3 is chosen to yield null, moderate, and strong associations between outcome Y and mediator M and covariates X, as detailed in the Appendix.

Table 2.1 shows that both proposed efficient estimators (LE and RE) have nearly identical performance. Relative to the RMLE, which is based on linear main effect models for outcome and mediator, they deliver drastic efficiency gains for the natural indirect effect, although not for the natural direct effect (see Table 2.15 in the Appendix). This is not surprising, considering the discussion in Section 2.2.3. In particular, the expressions for the direct and indirect effect in that section are based on two regressions: (a) a regression of mediator on treatment, which may or may not include the baseline covariates; and (b) a regression of outcome on treatment and mediator, which must include the baseline covariates when they confound the association between mediator and outcome. Model (a) is only needed for the estimation of the indirect effect. The choice whether or not to include baseline covariates may thus only affect estimation of the indirect effect; as stated in Section 2.2.3, adjustment is indeed favourable, in line with the properties of ordinary least squares estimation.

Association		RMLE	LE	RE	MLE	MLE$_\text{I}$	TMLE$_\text{P}$	TMLE$_\text{PP}$	TMLE
$Y \sim M$: moderate	Bias	0.015	0.006	0.006	0.004	0.006	0.020	0.020	0.020
$Y \sim X$: moderate	Emp SD	0.195	0.156	0.156	0.131	0.156	0.172	0.172	0.171
$M \sim X$: null	Mean SE	0.197	0.157	0.157	0.127	0.157	0.172	0.884	0.883
	Coverage	0.96	0.94	0.94	0.95	0.94	0.95	1.00	1.00
$Y \sim M$: moderate	Bias	0.007	0.006	0.006	0.004	0.006	0.009	0.009	0.006
$Y \sim X$: moderate	Emp SD	0.225	0.156	0.156	0.131	0.156	0.206	0.206	0.202
$M \sim X$: strong	Mean SE	0.228	0.157	0.157	0.127	0.157	0.206	0.973	0.973
	Coverage	0.95	0.94	0.94	0.95	0.94	0.95	1.00	1.00
$Y \sim M$: null	Bias	0.012	0.003	0.003	0.001	0.003	0.010	0.010	0.010
$Y \sim X$: moderate	Emp SD	0.174	0.128	0.128	0.089	0.128	0.147	0.147	0.137
$M \sim X$: moderate	Mean SE	0.176	0.130	0.130	0.090	0.130	0.151	0.850	0.850
	Coverage	0.96	0.96	0.96	0.95	0.95	0.96	1.00	1.00
$Y \sim M$: strong	Bias	0.009	0.009	0.009	0.007	0.009	0.020	0.020	0.019
$Y \sim X$: moderate	Emp SD	0.292	0.225	0.225	0.212	0.225	0.278	0.279	0.277
$M \sim X$: moderate	Mean SE	0.294	0.220	0.220	0.201	0.220	0.277	1.175	1.175
	Coverage	0.95	0.94	0.94	0.93	0.94	0.95	1.00	1.00
$Y \sim M$: moderate	Bias	0.002	0.006	0.006	0.004	0.006	0.006	0.006	0.003
$Y \sim X$: null	Emp SD	0.190	0.156	0.156	0.131	0.156	0.165	0.165	0.150
$M \sim X$: moderate	Mean SE	0.190	0.157	0.157	0.127	0.157	0.164	0.861	0.861
	Coverage	0.95	0.94	0.94	0.95	0.94	0.96	1.00	1.00
$Y \sim M$: moderate	Bias	0.014	0.006	0.006	0.004	0.006	0.018	0.018	0.016
$Y \sim X$: strong	Emp SD	0.222	0.156	0.156	0.131	0.156	0.202	0.202	0.200
$M \sim X$: moderate	Mean SE	0.225	0.157	0.157	0.127	0.157	0.203	0.967	0.967
	Coverage	0.97	0.94	0.94	0.95	0.94	0.95	1.00	1.00

Table 2.1: Simulation results for indirect effect on a continuous outcome under correct model specification.

The simulation results moreover show that the proposed efficient estimators attain nearly the same efficiency as the MLE without treatment interactions. The relatively minor efficiency benefit of this MLE comes exclusively from the additional assumption that the mediator and covariate effects on the outcome do not interact with treatment. The proposed estimators, LE and RE, do not make this assumption as they postulate separate models for the expected outcome in the treated and untreated populations, although one may equally well choose to fit more restrictive models. In fact, when that same assumption is made in the LE-estimator, then it reduces to the MLE; this follows from the remark at the end of Section 2.2.4

that the MLE is robust to misspecification of the mediator model when all models are linear, in which case it reduces to the LE estimator. When the MLE is based on an outcome model which includes treatment interactions, then it loses its relatively minor efficiency benefit over the LE an RE-estimators. Unlike for the direct effect, worse results in terms of efficiency are generally obtained for the TMLE. This is not the result of using a more flexible model for the outcome, considering that a similar efficiency loss is observed for the ‘parametric’ and ‘partly parametric’ TMLE estimators, but because these estimators do not presume correct specification of the outcome model to increase efficiency (i.e. they increase efficiency relative to the nonparametric model, rather than the model that assumes a correctly specified outcome mean). All competing estimators are unbiased and confidence intervals reach the nominal level, except for the ‘partly parametric’ and ‘non-parametric’ TMLE. This is because their sandwich standard errors, which are calculated as the variance of the influence function under the assumption of known nuisance parameters (see the Appendix), can be severely biased (under misspecification of models for the outcome and treatment).

It follows from the theoretical results in the Appendix that the potential of the proposed estimators to improve efficiency depends on the strength of the associations between outcome and mediator, outcome and covariates, and mediator and covariates. Not surprisingly, bigger efficiency gains can be realised when covariates X become more strongly predictive of outcome Y. The same is observed when the strength of the association between covariates X and mediator M is changed, although there may be scenarios where this is less obvious. While a stronger association between covariates and mediator improves estimation of the treatment effect on the mediator, it can also deteriorate the estimation of the mediator’s effect on the outcome as a result of multicollinearity. When the strength of the association between the mediator M and the outcome Y is changed, a stronger reduction in absolute terms is observed. This can be intuitively understood by considering the product-of-coefficient estimator of the indirect effect. Its Delta method standard error weighs the variance of the estimated treatment effect on the mediator by the magnitude of the mediator’s effect on the outcome. Since covariate adjustment reduces this variance, the extent to which it reduces is larger (in absolute terms) as the mediator is more strongly predictive of the outcome.

2.3.1.2 Binary Outcome

Association		RMLE	LE	RE	MLE	MLE$_\text{I}$	TMLE$_\text{P}$	TMLE$_\text{PP}$	TMLE
$Y \sim M$: moderate	Bias	0.002	0.002	0.002	0.139	0.002	0.001	0.001	0.045
$Y \sim X$: moderate	Emp SD	0.029	0.027	0.027	0.024	0.027	0.029	0.029	0.032
$M \sim X$: null	Mean SE	0.030	0.029	0.029	0.023	0.028	0.032	0.063	0.060
	Coverage	0.96	0.96	0.96	0.00	0.96	0.97	1.00	0.98
$Y \sim M$: moderate	Bias	-0.001	0.000	0.000	0.103	0.000	-0.001	-0.001	0.010
$Y \sim X$: moderate	Emp SD	0.033	0.025	0.026	0.024	0.025	0.033	0.033	0.030
$M \sim X$: strong	Mean SE	0.034	0.025	0.026	0.023	0.025	0.035	0.068	0.066
	Coverage	0.96	0.96	0.96	0.01	0.95	0.96	1.00	1.00
$Y \sim M$: null	Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$Y \sim X$: moderate	Emp SD	0.031	0.030	0.030	0.021	0.030	0.031	0.031	0.019
$M \sim X$: moderate	Mean SE	0.031	0.030	0.030	0.021	0.029	0.033	0.072	0.072
	Coverage	0.95	0.94	0.94	0.94	0.94	0.95	1.00	1.00
$Y \sim M$: strong	Bias	0.000	0.001	0.001	0.143	0.001	0.000	0.000	0.017
$Y \sim X$: moderate	Emp SD	0.033	0.027	0.027	0.024	0.026	0.033	0.033	0.031
$M \sim X$: moderate	Mean SE	0.033	0.027	0.027	0.024	0.026	0.035	0.064	0.063
	Coverage	0.96	0.96	0.96	0.00	0.95	0.96	1.00	1.00
$Y \sim M$: moderate	Bias	0.000	0.001	0.001	0.124	0.001	0.000	0.000	0.021
$Y \sim X$: null	Emp SD	0.031	0.027	0.027	0.024	0.027	0.031	0.031	0.029
$M \sim X$: moderate	Mean SE	0.032	0.027	0.028	0.023	0.027	0.033	0.065	0.063
	Coverage	0.96	0.96	0.95	0.00	0.95	0.97	1.00	1.00
$Y \sim M$: moderate	Bias	0.001	0.001	0.001	0.113	0.001	0.000	0.000	0.017
$Y \sim X$: strong	Emp SD	0.033	0.027	0.027	0.023	0.026	0.033	0.033	0.031
$M \sim X$: moderate	Mean SE	0.032	0.026	0.026	0.023	0.025	0.034	0.067	0.065
	Coverage	0.95	0.95	0.94	0.00	0.94	0.96	1.00	1.00

Table 2.2: Simulation results for indirect effect on a binary outcome under correct model specification.

Covariates $X = (X_1, ..., X_8)^t$ and mediator M are generated as before. A binary outcome Y is generated as a Bernoulli variate obeying logit$\{P(Y = 1|A = a, M, X)\} = \alpha_{0a} + \alpha_{1a}M + \alpha_{2a}^t X$ with $a = 0$ or 1, $(\alpha_{00}, \alpha_{01})$ equal to $(-0.8, 0.8)$, α_{1a} equal to $(0.6, -0.8)$ and α_{2a}^t chosen to yield null, moderate, and strong associations between outcome Y and mediator M and covariates X, as detailed in the Appendix. The simulation results in Table 2.2, as well as Table 2.16 of the Appendix, suggest similar conclusions as for the continuous outcome: efficiency gains relative to the RMLE and TMLE. In contrast to the results of the continuous outcome, the ‘non-parametric’ TMLE, based on a more flexible outcome model, shows some bias for the binary outcome. As before, we observe a relatively minor efficiency benefit of the MLE without treatment interactions in the outcome model, but also a large bias since the data generating mechanism included interactions in the binary outcome model. Both, the minor efficiency gain and the large bias, disappear when treatment interactions are included in the outcome model.

2.3.2 Misspecification of the model for the mediator

A drawback of direct maximum likelihood estimation based on the mediation formula is that it requires correct specification of the model for the mediator distribution. The proposed estimators also make use of such a model, but only to increase efficiency; its misspecification does not induce bias. In this section, we will examine this through various simulation settings with misspecification of the mediator model, chosen to reflect important, but realistic degrees of misspecification. Covariates $X^* = (X_1, ..., X_8)^t$ are generated as before. The continuous mediator M is drawn from a normal distribution with mean $E(M|A,X) = \beta_0 + \beta_1 A + \beta_2^t X$. In the first scenario, the mediator was drawn from a normal distribution with residual variance 1, and with X including the eight covariates X^* and two squared terms X_1^2 and X_2^2. Parameter values were set to $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = (1,0,0,0,0,0,0,0,-1,-1)^t$. As a result, mediator models in the analyses are misspecified since the two squared terms were ignored. In the second setting, M has mean $E(M|A,X) = \beta_0 + \beta_1 A + \beta_2^t X + \beta_3 U_n$ and residual variance 1, where U_n is Bernoulli distributed with $P(U_n = 1) = 0.5$ and parameter values $\beta_0 = 0$, $\beta_1 = -1.25$, $\beta_2 = (-0.25, 0.25, 0, 0, 0, 0, 0, 0)^t$ and $\beta_3 = 10$. Mediator models are misspecified because they did not include U_n as a predictor. In the third scenario, X equals X^* with $\beta_0 = 0$, $\beta_1 = 0.5$, and $\beta_2 = (0.1, -0.2, 0.1, 0.2, 0.15, -0.2, -0.15, -0.05)^t$, but residual errors were taken from a Student t-distribution with three degrees of freedom, multiplied by 1. In the fourth and final setting, the continuous mediator M is drawn from a gamma distribution with shape equal to $\beta_0 + \beta_1 A + \beta_2^t X$ with $\beta_0 = 1.5$, $\beta_1 = 1.5$, $\beta_2 = (0.2, 0, 0, 0, 0, 0, 0, 0)^t$ and scale equal to 1; here, X equals X^*, but $X_1 \sim \mathcal{N}(6,1)$.

2.3.2.1 Continuous Outcome

The continuous outcome Y is drawn from a normal distribution with mean $E(Y|A,M,X) = \alpha_0 + \alpha_1 A + \alpha_2 M + \alpha_3^t X$ and residual variance 4, and the following parameter values: $\alpha_0 = 1$, $\alpha_1 = 2$, $\alpha_2 = -1$ and $\alpha_3 = (0.8, 0.7, 0.55, -0.6, -0.25, 0, 0, 0)^t$.

Misspecification		RMLE	LE	RE	MLE	MLE$_\text{I}$	TMLE$_\text{P}$	TMLE$_\text{PP}$	TMLE
Not included higher order terms	Bias	0.007	0.001	0.001	0.001	0.001	-0.001	-0.001	0.002
	Emp SD	0.230	0.213	0.212	0.209	0.212	0.230	0.230	0.229
	Mean SE	0.231	0.211	0.211	0.205	0.210	0.230	0.913	0.913
	Coverage	0.95	0.95	0.95	0.94	0.95	0.94	1.00	1.00
Forgotten predictor	Bias	0.004	-0.001	-0.001	0.000	-0.001	-0.093	-0.094	-0.092
	Emp SD	0.477	0.472	0.471	0.471	0.471	0.479	0.479	0.479
	Mean SE	0.474	0.454	0.454	0.457	0.456	0.474	1.434	1.432
	Coverage	0.95	0.94	0.94	0.93	0.94	0.95	1.00	1.00
Outliers in mediator distribution	Bias	-0.002	-0.012	-0.012	-0.013	-0.012	-0.015	-0.015	-0.016
	Emp SD	0.214	0.163	0.163	0.161	0.163	0.212	0.212	0.211
	Mean SE	0.207	0.160	0.160	0.156	0.159	0.204	1.330	1.330
	Coverage	0.95	0.94	0.94	0.94	0.94	0.94	1.00	1.00
Gamma mediator distribution	Bias	0.002	-0.008	-0.008	-0.005	-0.008	0.017	0.017	0.017
	Emp SD	0.220	0.183	0.183	0.171	0.183	0.207	0.207	0.207
	Mean SE	0.226	0.189	0.189	0.177	0.188	0.215	1.109	1.109
	Coverage	0.95	0.95	0.95	0.96	0.95	0.96	1.00	1.00

Table 2.3: Indirect effect on a continuous outcome with mediator misspecification

Table 2.3, as well as Table 2.17 of the Appendix summarise the simulation results for the continuous outcome under mediator model misspecification. Both proposed estimators are again more efficient than the restricted MLE and the TMLE in terms of the natural indirect effect estimate, even if the model for the mediator is misspecified. The MLE without treatment interaction continues to be unbiased because the linear model for the mediator satisfies identity 2.10, indicating that the MLE is robust against mediator model misspecification. It is also slightly more efficient, which reflects its reliance on a more restrictive outcome model, which does not allow for modification of the mediator and covariate effects by treatment. If such mediator and covariate treatment interactions are included, the slight efficiency gain of the MLE disappears.

2.3.2.2 Binary Outcome

The binary outcome Y is generated as Bernoulli according to $\text{logit}\{P(Y=1|A=a,M,X)\}=\alpha_{0a}+\alpha_{1a}M+\alpha^t_{2a}X$ with $a=0$ or 1. Parameter values are detailed in the Appendix.

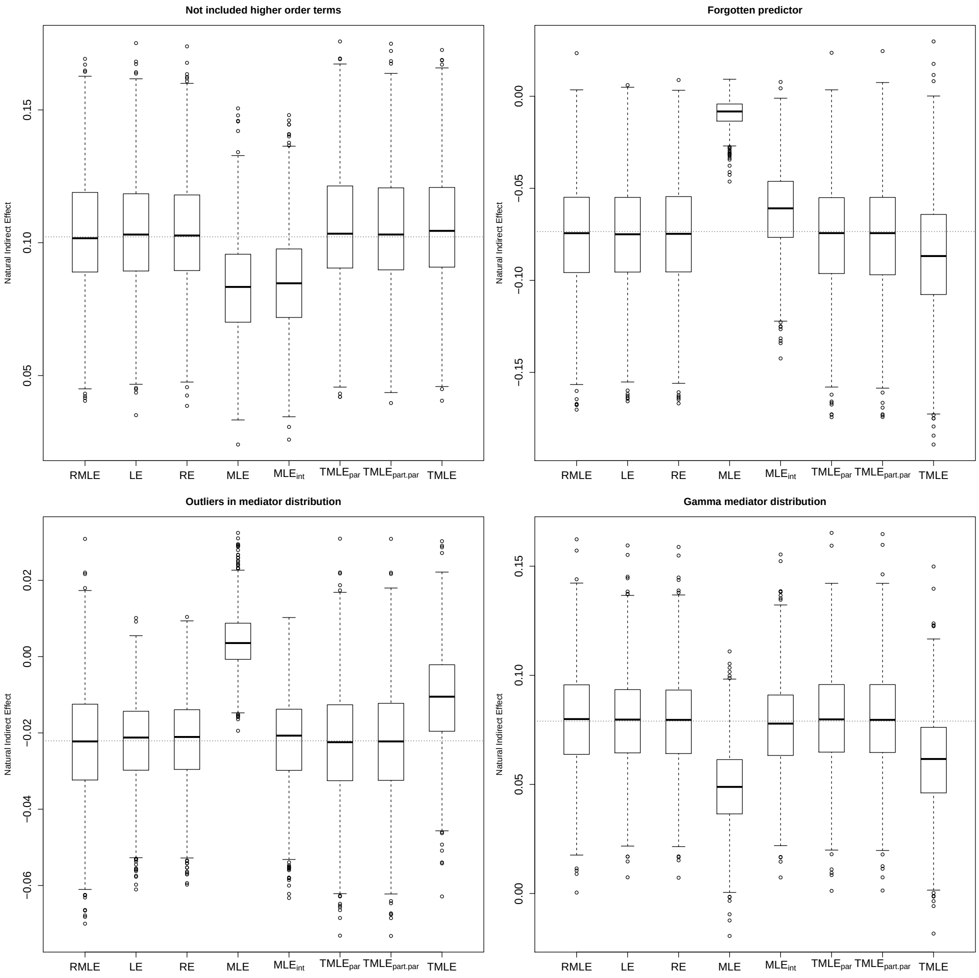


Figure 2.1: Indirect effect on a binary outcome with mediator misspecification

Table 2.4 and Figure 2.1 (and Table 2.18 and Figure 2.2 of the Appendix) suggest similar conclusions as before in terms of efficiency. The MLE is once again extremely biased when it does not take into account that the mediator and covariate effects on the outcome interact with treatment, which leads to undercoverage of the confidence intervals. Even when modelling those interactions, forgetting to include an important predictor (see also Table 2.18 and Figure 2.2 of the Appendix) or not including higher order terms induces some bias and leads to undercoverage. In three of the four settings (i.e. forgotten predictor, outliers and gamma distribution), the 'non-parametric' TMLE seems more vulnerable to bias with a binary outcome.

Misspecification		RMLE	LE	RE	MLE	MLE_I	TMLE_P	TMLE_PP	TMLE
Not included higher order terms	Bias	0.000	0.001	0.000	-0.020	-0.018	0.002	0.002	0.003
	Emp SD	0.022	0.021	0.021	0.019	0.020	0.022	0.022	0.022
	Mean SE	0.021	0.020	0.020	0.019	0.019	0.022	0.024	0.023
	Coverage	0.94	0.94	0.94	0.82	0.84	0.94	0.96	0.96
Forgotten predictor	Bias	-0.002	-0.002	-0.002	0.064	0.012	-0.002	-0.002	-0.013
	Emp SD	0.030	0.030	0.030	0.007	0.024	0.031	0.031	0.034
	Mean SE	0.030	0.029	0.029	0.007	0.023	0.032	0.070	0.068
	Coverage	0.95	0.94	0.94	0.00	0.90	0.96	1.00	1.00
Outliers in mediator distribution	Bias	-0.001	0.000	0.000	0.026	0.000	-0.001	-0.001	0.011
	Emp SD	0.015	0.012	0.012	0.008	0.012	0.015	0.015	0.014
	Mean SE	0.017	0.014	0.014	0.008	0.013	0.019	0.060	0.059
	Coverage	0.97	0.98	0.98	0.11	0.95	0.98	1.00	1.00
Gamma mediator distribution	Bias	-0.001	-0.001	-0.001	-0.031	-0.003	0.000	0.000	-0.020
	Emp SD	0.023	0.022	0.022	0.019	0.021	0.023	0.024	0.024
	Mean SE	0.024	0.023	0.023	0.019	0.022	0.026	0.038	0.039
	Coverage	0.96	0.96	0.96	0.62	0.94	0.97	1.00	0.99

Table 2.4: Indirect effect on a binary outcome with mediator misspecification

2.3.3 Additional results

In the Appendix, we provide a large number of additional simulation results on the behaviour of these estimators in small samples and under misspecification of the outcome model. With outcome model misspecification, the TMLE performed sometimes slightly better and sometimes slightly worse in terms of bias and standard deviation than the considered estimators. Overall, the results are largely similar as before.

2.4 Data analysis

We re-analyze data from a psychological experiment on the effect of implicit priming with social deception (A) on responses towards other's pain (Y) (De Ruddere et al. 2013). In total, 55 participants were randomly assigned to the neutral condition ($n = 27$) or the social deception condition ($n = 28$), in which they read either a neutral text about the use of the health care system or a text about its misuse, respectively. It was hypothesised that implicit priming with social deception would lower the observers' estimates of pain experienced by a patient (presented to the participant in a video). A numerical rating scale, going from 0 to 10, was used to asses observers' estimated pain of the patient, where 10 indicated 'pain as bad as could be'. Furthermore, the patients' valence (M), i.e. how positive or negative a patient was evaluated by the observers (based a 21-point scale from -10 = very negative to 10 = very positive), was considered as a potential mediator. It was explored whether social deception induced less positive ratings, which in turn lowered the observers' estimates of pain experienced by a patient. Available baseline covariates (C) include education, profession, gender, marital status and age.

	Estimator	Estimate	SE_{boot}	$\text{SE}_{\text{sandwich}}$	95% CI
Direct effect	RMLE	0.49	0.39	0.28	-0.27 to 1.25
	LE	0.54	0.38	0.26	-0.20 to 1.28
	RE	0.51	0.33	0.26	-0.14 to 1.16
	MLE	0.40	0.29		-0.17 to 0.97
	TMLE_{PP}	0.41	0.29	0.20	-0.16 to 0.98
Indirect effect	RMLE	-0.58	0.38	0.26	-1.32 to 0.16
	LE	-0.22	0.25	0.14	-0.71 to 0.27
	RE	-0.21	0.22	0.14	-0.64 to 0.22
	MLE	-0.11	0.11		-0.33 to 0.11
	TMLE_{PP}	-0.52	0.27	1.00	-1.05 to 0.01

Table 2.5: Estimates of the natural direct and indirect effect without model building.

In a first stage, we used main effect models that include all baseline factors and mean-centered age. The corresponding estimated natural direct and indirect effects are presented in Table 2.5. We did not observe a significant direct effect of priming on the rating of patients' pain, nor did we observe a significant indirect effect via the evaluation of the patients' valance. The 'partially parametric' TMLE estimator of Lendle et al. (2013), based on the same outcome model as the MLE, showed a slightly larger bootstrap bias (respectively -0.02 and 0.06 for the direct and indirect effect in comparison to for example -0.01 and 0.01 of the LE estimator), smaller standard errors for the direct effect and larger ones for the indirect effect. The discrepancy between the bootstrap standard errors and the sandwich standard errors of all other estimators suggests that the latter might have also been underestimated because of the small sample. Similar to the simulation results, we observe considerably smaller standard errors for the LE-, RE- and TMLE-estimator of the indirect effect than for the RMLE-estimator.

Next, the models were extended to only include important main and interaction effects. A forward selection procedure with an inclusion criterion of 0.15 for the p-value was used. Tables 2.19, 2.20, 2.21, 2.22 and 2.23 in the Appendix show the fitted outcome models for $E(Y|A=1,M,X)$ and $E(Y|A=0,M,X)$, for $E(Y|A=1,X)$ and $E(Y|A=0,X)$, and for $E\{m(M,X;\eta^*)|A=0,X\}$, respectively. We additionally evaluated a TMLE estimator that makes use of data-adaptive learning algorithms for the outcome mean (2.7). Bootstrap standard errors were evaluated conditional on the selected model. Overall, the obtained estimates are slightly more efficient than before, but deliver similar conclusions. One exception is that we now observe a significant direct effect of priming on the rating of patients' pain of 0.63 (95% CI 0.003 to 1.26) based on the RE estimator. We conclude that if a participant were primed with social deception, but his or her evaluation of the patient's valence stayed as it would have been without the priming, the participants' score of the patients' pain would on average be 0.63 points higher than in the neutral condition.

The 'partially parametric' TMLE resulted in a significant indirect effect of priming via patients' valence of -0.63 (95% CI -1.12 to -0.14). It appears favourable in terms of efficiency, but has a relatively large bootstrap bias of 0.12. The reason for this increased efficiency may be the fact that the outcome model for the 'partially parametric' TMLE is fitted on the whole sample, while the LE and RE estimators postulate separate models for the expected outcome in the treated and untreated populations. Although the LE and RE estimators can equally be based on more restrictive models, we generally advise to model the treated and untreated populations separately in view of possible model misspecification. Similarly, the TMLE estimator based on data-adaptive learning algorithms showed a larger bootstrap bias (respectively -0.10 and 0.13 for the direct and indirect effect in comparison to for example -0.04 and 0.04 of the LE estimator), which is likely the result of the small sample size.

	Estimator	Estimate	SE_{boot}	$\text{SE}_{\text{sandwich}}$	95% CI
Direct effect	RMLE	0.51	0.36	0.29	−0.20 to 1.22
	LE	0.63	0.33	0.27	−0.02 to 1.28
	RE	0.63	0.32	0.27	0.003 to 1.26
	MLE	0.41	0.28		−0.14 to 0.95
	TMLE_{PP}	0.41	0.28	0.20	−0.14 to 0.96
	TMLE	0.19	0.29	0.21	−0.38 to 0.76
Indirect effect	RMLE	−0.60	0.33	0.25	−1.25 to 0.05
	LE	−0.44	0.29	0.22	−1.01 to 0.13
	RE	−0.43	0.29	0.22	−1.00 to 0.14
	MLE	−0.14	0.10		−0.37 to 0.04
	TMLE_{PP}	−0.63	0.25	0.98	−1.12 to −0.14
	TMLE	−0.46	0.29	0.92	−1.03 to 0.11

Table 2.6: Estimates of the natural direct and indirect effect with extended models.

2.5 Discussion

In this article, we have proposed estimators of the natural direct and indirect effect of a randomised treatment on an end-of-study outcome, with respect to a given mediator. Like popular maximum likelihood estimators (MLE) based on the mediation formula, our estimators make use of a model for the mediator to improve efficiency, but unlike MLEs, they are robust to its misspecification. In particular, we have considered two estimators that optimally extract information from the baseline covariate data: one (LE) that achieves efficiency when a working model for the mediator is correctly specified, and one (RE) that achieves efficiency within a more restrictive class of estimators regardless of correct specification of a working model for the mediator. In simulation studies, we found both estimators to exhibit similar behaviour and therefore recommend the LE estimator for practical use in view of its greater simplicity. A by-product of our results is that it provides conditions under which the MLE is robust against mediator model misspecification. All results extend directly to settings with multiple mediators, upon redefining M to be the vector of mediators.

Our proposal extends to the analysis of observational studies, where the treatment A is not randomly assigned, but the covariate set X remains sufficient to adjust for confounding of the effects of exposure on outcome and of exposure on mediator. Because the propensity scores $P(A=1|X)$ are generally unknown in such studies, consistent estimates are then required. When the covariate set X is low-dimensional, then the propensity scores can be estimated non-parametrically. In that case, the proposed locally efficient estimator is asymptotically equivalent to the restricted MLE because its potential to exploit covariate information is predicated on the existence of residual covariate imbalances between the exposure groups. When the covariate set X is high-dimensional, then an additional model for the propensity score is required. It then follows from Tsiatis et al. (2008) that the proposed estimator for $E\left(Y_0\right)$ is double-robust: unbiased in large samples when either the model for the propensity score (i.e. working model 1) or the model for the outcome in the untreated (i.e. working model 2) is correctly specified, but not necessarily both. It is immediate from the general expressions (2.13) and (2.14) in the Appendix, that a similar result holds for the proposed estimator for $E\left(Y_{1M_0}\right)$. It is double-robust in the following sense: unbiased in large samples when either the model for the propensity score (i.e. working model 1) or the model for the fitted outcome in the untreated (i.e. working model 3) is correctly specified, but not necessarily both. Our proposal is closely related to the triple-robust estimators of Tchetgen Tchetgen and Shpitser (2012), which are unbiased in large samples when two out of three models for the exposure, mediator and outcome are correctly specified, and to related work on Targeted Maximum Likelihood estimation (Lendle et al. 2013), which we evaluated in our simulation studies. It is also closely related to the double-robust estimators of Vansteelandt et al. (2012a), which are unbiased in large samples when either the outcome model, or both models for the exposure and mediator are correctly specified. Our proposal differs from these other proposals in that it makes explicit use of the known randomisation probabilities; it in particular avoids the need for inverse weighting by the mediator distribution.

Like other mediation analyses, our analysis of natural direct and indirect effects relies on a technical requirement that the data have been generated by a so-called nonparametric structural equation model. Mediation analysis based on natural direct and indirect effects has therefore been the subject of recent debate (Robins and Richardson 2011; Naimi et al. 2014). We agree that the requirement of a nonparametric structural equation model is strong, but tend not be to overly concerned about it so long as one can rule out intermediate confounding. The reason is that in that case, the natural direct effects that we calculate remain interpretable as direct effects - even under violation of the nonparametric structural equation model - in the following sense (Petersen et al. 2006). They can then be interpreted as the (controlled) treatment effect that would be seen if for each individual, the mediator were fixed at some value randomly drawn from the conditional distribution of the mediator, given covariates X, in the controls. We refer the reader to Vansteelandt and VanderWeele (2013), VanderWeele et al. (2014) or Tchetgen Tchetgen and VanderWeele (2014) for strategies to deal with intermediate confounders.

2.A Appendix

2.A.1 Identification assumptions

It follows from Pearl (2001) that the natural direct and indirect effects defined in Section 2.2.1 can be identified when $Y_a \perp\!\!\!\perp A|X$, $M_a \perp\!\!\!\perp A|X$, $Y_{am} \perp\!\!\!\perp M|A = a, X$ and $Y_{am} \perp\!\!\!\perp M_{a^*}|X$. It follows from VanderWeele and Vansteelandt (2009) that when the exposure A is randomly assigned conditional on X, then these assumptions are satisfied when the data have been generated by a so-called nonparametric structural equation model, A and X are sufficient to control for confounding of the association between M and Y, and none of the variables X is affected by the exposure.

2.A.2 General theory: known nuisance parameters

It follows from Tchetgen Tchetgen and Shpitser (2012) and Albert (2012) that

$$E(Y_{1M_0}) = E\left\{\frac{1-A}{P(A=0|X)}E(Y|A=1,M,X)\right\},$$

when the exposure is randomly assigned and X is sufficient to adjust for confounding of the association between M and Y (see Pearl (2001) and VanderWeele and Vansteelandt (2009) for further discussion of the required identification assumptions).

Suppose first that the randomisation probabilities (and thus α^*) are known, and that also η^* is known. Then all consistent and asymptotically normal estimators of $\theta = E(Y_{1M_0})$ can be obtained by solving an estimating equation of the form

$$\begin{aligned} 0 = \sum_{i=1}^{n} U_i(\theta,\alpha^*,\eta^*) &= \sum_{i=1}^{n} \frac{1-A_i}{1-g(X_i;\alpha^*)} m(M_i,X_i;\eta^*) - \theta + d(X_i)\left\{A_i - g(X_i;\alpha^*)\right\} \\ &\quad + e(M_i,X_i)A_i\left\{Y_i - m(M_i,X_i;\eta^*)\right\}, \end{aligned}$$

for some index function $d(X_i)$ and $e(M_i,X_i)$. The variance of the solution $\hat{\theta}$ to this equation equals 1 over n times the variance of $U_i(\theta,\alpha^*,\eta^*)$. The optimal choices of index functions $d(X_i)$ and $e(M_i,X_i)$ may thus be obtained by minimising the variance of $U_i(\theta,\alpha^*,\eta^*)$ w.r.t. $d(X_i)$ and $e(M_i,X_i)$. Since the terms $d(X_i)\{A_i - g(X_i;\alpha^*)\}$ and $e(M_i,X_i)A_i\{Y_i - m(M_i,X_i;\eta^*)\}$ are orthogonal (i.e., uncorrelated), the optimal choice of $d(X_i)$ may in particular be obtained by population least squares projection of $(1-A_i)m(M_i,X_i;\eta^*)/\{1-g(X_i;\alpha^*)\}$ onto $A_i - g(X_i;\alpha^*)$. It thus equals

$$\begin{aligned} d_{\text{opt}}(X_i) &= -\frac{E\left[\frac{1-A_i}{1-g(X_i;\alpha^*)}m(M_i,X_i;\eta^*)\{A_i-g(X_i;\alpha^*)\}|X_i\right]}{E\left[\{A_i-g(X_i;\alpha^*)\}^2|X_i\right]} \\ &= \frac{E\left[(1-A_i)\frac{g(X_i;\alpha^*)}{1-g(X_i;\alpha^*)}m(M_i,X_i;\eta^*)|X_i\right]}{g(X_i;\alpha^*)\{1-g(X_i;\alpha^*)\}} \\ &= \frac{E\left[(1-A_i)m(M_i,X_i;\eta^*)|X_i\right]}{\{1-g(X_i;\alpha^*)\}^2} \\ &= \frac{E\left[m(M_i,X_i;\eta^*)|A_i=0,X_i\right]}{1-g(X_i;\alpha^*)}. \end{aligned}$$

Likewise, the optimal choice of $e(M_i,X_i)$ can be obtained by population least squares projection of $(1-A_i)m(M_i,X_i;\eta^*)/\{1-g(X_i;\alpha^*)\}$ onto $A_i\{Y_i - m(M_i,X_i;\eta)\}$. It equals 0.

Calculating the efficient estimator of $E(Y_{1M_0})$ thus requires a working model for the conditional expectation $E\left[m(M_i,X_i;\eta^*)|A_i=0,X_i\right]$, which we here more generally formalise as

$$E\left[m(M_i,X_i;\eta^*)|A_i=0,X_i\right] = z(X_i;\gamma^*),$$

where $z(X;\gamma)$ is a known function, smooth in γ and γ^* is an unknown finite-dimensional parameter. For given estimator $\hat{\gamma}$ of γ^*, the efficient estimator is then obtained as

$$\begin{aligned} \hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1-g(X_i;\alpha^*)}\left[(1-A_i)m(M_i,X_i;\eta^*)+z(X_i;\hat{\gamma})\{A_i-g(X_i;\alpha^*)\}\right] \quad (2.13) \\ &= \frac{1}{n}\sum_{i=1}^{n}z(X_i;\hat{\gamma})+\frac{1-A_i}{1-g(X_i;\alpha^*)}\left[m(M_i,X_i;\eta^*)-z(X_i;\hat{\gamma})\right]. \quad (2.14) \end{aligned}$$

Its variance is straightforwardly obtained as the variance of the corresponding sample average, ignoring estimation of $\hat{\gamma}$.

The first term in the expression for $\hat{\theta}$,

$$\frac{1}{n}\sum_{i=1}^{n} z(X_i;\hat{\gamma}) \tag{2.15}$$

is closely related to the mediation formula. It involves averaging the expected outcome values, if the exposure were set to 1, over the mediator distribution if the exposure were set to zero. The second contribution insulates it against bias resulting from possible misspecification of the mediator distribution.

If $z(X_i;\gamma^*)$ is a canonical generalised linear model that includes an intercept and γ^* is estimated using a standard maximum likelihood procedure on data for the unexposed, using weights $1/\{1-g(X_i;\alpha^*)\}$, then the efficient estimator reduces to (2.15) and thus becomes a so-called substitution estimator. This can be seen because the remaining contribution then equals the score of the intercept in the fitted model, which is set to zero through the fitting procedure for γ^*.

Finally, while misspecification of the working model for the conditional expectation $E\left[m(M_i,X_i;\eta^*)|A_i=0,X_i\right]$ does not affect the consistency of $\hat{\theta}$, it does affect its efficiency and may in particular make the estimator less efficient than the restricted MLE that would be obtained upon setting $d(X_i)=0$. When α^* is known, this can be remedied by letting $z(.)$ be the identity link and then estimating γ^* via ordinary least squares regression of $(1-A)m(M,X;\eta^*)/\{1-g(X;\alpha^*)\}$ onto $(1,X)\{A-g(X;\alpha^*)\}/\{1-g(X;\alpha^*)\}$. This is guaranteed to increase efficiency because the effect of least squares projection is to minimise sums of squares (and thus variance). Likewise, for estimation of $E(Y_0)$, one may consider estimating β^* via ordinary least squares regression of $(1-A)Y/\{1-g(X;\alpha^*)\}$ onto $(1,X)\{A-g(X;\alpha^*)\}/\{1-g(X;\alpha^*)\}$. When α^* is unknown, then we adopt the same principle after projecting the influence function $(1-A)m(M,X;\eta^*)/\{1-g(X;\alpha^*)\}-\gamma^{t}X\{A-g(X;\alpha^*)\}/\{1-g(X;\alpha^*)\}$ onto the orthocomplement of the tangent space for α^*, which we show for the case $g(X;\alpha^*)=\alpha^*$:

$$\frac{(1-A)m(M,X;\eta^*)}{(1-\alpha^*)}-\gamma^{t}X\frac{(A-\alpha^*)}{(1-\alpha^*)}+E\left[\frac{(1-A)m(M,X;\eta^*)}{(1-\alpha^*)^2}+\frac{\gamma^{t}X}{(1-\alpha^*)}\right](A-\alpha^*)$$

$$\frac{(1-A)m(M,X;\eta^*)+E\{m(M,X;\eta^*)|A=0\}(A-\alpha^*)}{(1-\alpha^*)}-\gamma'\{X-E(X)\}\frac{(A-\alpha^*)}{(1-\alpha^*)},$$

which suggestes estimating γ^* via ordinary least squares regression of $(1-A)m(M,X;\eta^*)+E\{m(M,X;\eta^*)|A=0\}(A-\alpha^*)$ onto $\{X-E(X)\}(A-\alpha^*)$.

2.A.3 General theory: unknown nuisance parameters

Suppose now that η^* is unknown but substituted by a consistent estimator $\hat{\eta}$, obtained by solving an estimating equation of the form

$$0=\sum_{i=1}^{n}U_{\eta,i}(\eta^*)=\sum_{i=1}^{n}e_\eta(M_i,X_i)A_i\{Y_i-m(M_i,X_i;\eta^*)\}.$$

Then all consistent and asymptotically normal estimators of $E(Y_{1M_0})$ may still be obtained by solving an estimating equation of the above form. However, the variance of the solution $\hat{\theta}$ is now equal to 1 over n times the variance of

$$U_i^*(\theta,\alpha^*,\eta^*) \equiv U_i(\theta,\alpha^*,\eta^*)-E\left\{\frac{\partial}{\partial\eta}U_i(\theta,\alpha^*,\eta^*)\right\}E\left\{\frac{\partial}{\partial\eta}U_{\eta,i}(\eta^*)\right\}^{-1}U_{\eta,i}(\eta^*).$$

Since $U_{\eta,i}(\eta^*)$ is of the form $e_\eta(M_i,X_i)A_i\{Y_i-m(M_i,X_i;\eta^*)\}$, (2.16) has the same form of $U_i(\theta,\alpha^*,\eta^*)$ (although corresponding to a different choice of $e(M_i,X_i)$). The optimal choice of $d(X_i)$ that corresponds to an estimator of θ^* with minimal asymptotic variance is thus $d^{\text{opt}}(X_i)$, as before.

Suppose now that $g(X_i;\alpha)=\alpha$, where α^* is substituted by a consistent estimator $\hat{\alpha}$, obtained by solving an estimating equation of the form

$$0=\sum_{i=1}^{n}U_{\alpha,i}(\alpha^*)=\sum_{i=1}^{n}A_i-\alpha^*.$$

Then the variance of the solution $\hat{\theta}$ is equal to 1 over n times the variance of

$$\begin{aligned}U_i^{**}(\theta,\alpha^*,\eta^*) \equiv\ & U_i(\theta,\alpha^*,\eta^*)-E\left\{\frac{\partial}{\partial\eta}U_i(\theta,\alpha^*,\eta^*)\right\}E\left\{\frac{\partial}{\partial\eta}U_{\eta,i}(\eta^*)\right\}^{-1}U_{\eta,i}(\eta^*)\\ & -E\left\{\frac{\partial}{\partial\alpha}U_i(\theta,\alpha^*,\eta^*)\right\}E\left\{\frac{\partial}{\partial\alpha}U_{\alpha,i}(\alpha^*)\right\}^{-1}U_{\alpha,i}(\alpha^*).\end{aligned}$$

It follows by a similar reasoning as in Appendix D.2.2 of Bartlett et al. (2014) that this leaves the optimal choice $d^{\text{opt}}(X_i)$ unchanged. This is not surprising because, by construction, $U_i(\theta,\alpha^*,\eta^*)$ with $d(X_i)=d^{\text{opt}}(X_i)$ is orthogonal to the scores for α^*. Note that this is not true for the restricted efficient estimator defined in Section 2.2.5, for which the results of Section 2.2.5 must now be used!

Because by construction of $d^{\text{opt}}(X_i)$, $U_i^*(\theta,\alpha^*,\eta^*)$ is orthogonal to scores of the form $A_i-\alpha^*$, we further have that $E\left\{\frac{\partial}{\partial\alpha}U_i(\theta,\alpha^*,\eta^*)\right\}=0$, thus making adjustment for the estimation of α^* unnecessary, when the efficient choice $d^{\text{opt}}(X_i)$ is adopted. More generally, when α^* is estimated by maximum likelihood, then the influence function for θ becomes, up to a scalar,

$$\begin{aligned}&U_i^*(\theta,\alpha^*,\eta^*)-E\left\{\frac{\partial}{\partial\eta}U_i^*(\theta,\alpha^*,\eta^*)\right\}E\left\{\frac{\partial}{\partial\alpha}U_{\alpha,i}(\alpha^*)\right\}^{-1}U_{\alpha,i}\\&=U_i^*(\theta,\alpha^*,\eta^*)-E\left\{U_i^*(\theta,\alpha^*,\eta^*)S_{i,\alpha}\right\}E\left\{S_{i,\alpha}S_{i,\alpha}'\right\}^{-1}S_{i,\alpha};\end{aligned}$$

because the second term is a projection, we may conclude that efficient estimation of α^* (rather than using a known α^*) reduces variance.

2.A.4 Double-robust estimators

We refer the reader to Tchetgen Tchetgen and Shpitser (2012) and Lendle et al. (2013) for double-robust estimators of $E(Y_{1M_0})$: estimators that are unbiased in large samples when either model (2.7) or a model for the distribution of the mediator, given treatment and covariates (or instead of this, a model for the probability of treatment, given mediator and covariates) is correctly specified. In the simulation study, we followed the Targeted Maximum Likelihood Estimation (TMLE) approach of Lendle et al. (2013), which makes use of data-adaptive learning algorithms for the outcome mean (2.7) and for the probability of treatment, given mediator and covariates. We here explain how we estimated the indirect effect $E(Y_1-Y_{1M_0})$ in the simulation study, as this is not explained in Lendle et al. (2013). In particular, consider constant randomisation probabilities $g(X;\alpha)=\alpha$. Let $Q(1,M,X)$ be a working model for $E(Y|A=1,M,X)$ and $h(a,M,X)$ be a working model for $P(A=a|M,X)$ for $a=0,1$. Then it can be shown that the influence function of the indirect effect $\theta = E(Y_1 - Y_{1M_0})$ in the nonparametric model equals

$$\left\{\frac{A}{\alpha} - \frac{A}{1-\alpha}\frac{h(0,M,X)}{h(1,M,X)}\right\}\{Y - Q(1,M,X)\} + \{A - h(1,M,X)\} \times$$
$$\left(\frac{1}{\alpha} + \frac{1}{1-\alpha}\right)Q(1,M,X) + \left(\frac{h(1,M,X)}{\alpha} - \frac{h(0,M,X)}{1-\alpha}\right)Q(1,M,X) - \theta.$$

We started from initial fits $Q^{(0)}(1,M,X)$, $h^{(0)}(1,M,X)$ and $h^{(0)}(0,M,X) = 1 - h^{(0)}(1,M,X)$ obtained via the use of non-parametric data adaptive learning algorithms such as the super learner (van der Laan et al. 2007), which combines machine learning algorithms and parametric models using cross validation. GLM, Step, GLM interaction and Step interaction algorithms were used as prediction algorithms in the super learner function. Additionally, the parametric TMLE was fitted using the correct outcome model for initial fits of $Q^{(0)}(1,M,X)$. For a continuous outcome Y^*, we set $a = \min(Y^*)$, $b = \max(Y^*)$, and $Y = (Y^* - a)/(b-a)$. The initial estimates for $Q^{(0)}(1,M,X)$ of $E(Y^*|A = 1,M,X)$ and $h^{(0)}(1,M,X)$ of $P(A = 1|M,X)$ are represented as a logistic function of their logit transformation. Because logit(x) is not defined when $x = 0$ or 1, $Q^{(0)}(1,M,X)$ and $h^{(0)}(1,M,X)$ were bounded away from 0 and 1 by truncating at the α and $1-\alpha$ percentiles with $\alpha = 0.01$. In a next step, we iteratively update these estimates for $j = 1,2,...$ by fitting the parametric extensions until convergence:

$$\begin{aligned}\text{logit}Q^{(j)}(1,M,X) &= \text{logit}Q^{(j-1)}(1,M,X) + \varepsilon_y^{(j-1)}C_y^{(j-1)}\\ \text{logit}h^{(j)}(1,M,X) &= \text{logit}h^{(j-1)}(1,M,X) + \varepsilon_a^{(j-1)}C_a^{(j-1)},\end{aligned}$$

where

$$\begin{aligned}C_y^{(j-1)} &= \left\{\frac{A}{\hat{\alpha}} - \frac{A}{1-\hat{\alpha}}\frac{h^{(j-1)}(0,M,X)}{h^{(j-1)}(1,M,X)}\right\}\\ C_a^{(j-1)} &= \left(\frac{1}{\hat{\alpha}} + \frac{1}{1-\hat{\alpha}}\right)Q^{(j-1)}(1,M,X).\end{aligned}$$

and, upon convergence, calculate the estimator as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h^{(j)}(1,M_i,X_i)}{\hat{\alpha}}-\frac{h^{(j)}(0,M_i,X_i)}{1-\hat{\alpha}}\right)Q^{(j)}(1,M_i,X_i).$$

2.A.5 Simulation study

2.A.5.1 Correct model specification

Continuous outcome The scenario where there is no association between mediator M and covariates X is obtained, with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = (0,0,0,0,0,0,0,0)^T$ and residual variance $\sigma^2 = 1$. The scenario's with a moderate association between M and X are generated with $\beta_2 = (0.1,-0.2,0.8,0.15,0.2,-0.6,0.25,-0.5)^T$. In the setting with a strong association between M and X, β_2 is changed to $(0.1,-0.2,1.15,0.25,0.4,-1.3,0.5,-0.8)^T$. In terms of the strength of the association between outcome Y and covariates X, a setting with no association is created with $\alpha_0 = 1$, $\alpha_1 = 2$, $\alpha_3 = (0,0,0,0,0,0,0,0)^T$ and residual variance $\sigma^2 = 4$. In the moderate scenario, α_3 is modified to $(0.55,0.7,0.8,-0.55,-0.25,0,0,0)^T$ and in the strong setting to $(0.65,0.9,1.25,-1,-0.65,0,0,0)^T$. To vary the strength of the association between outcome Y and mediator M, α_2 was adjusted from 0 in the null scenario, to -1 in the moderate setting, and to -2 in the strong scenario.

Binary outcome Similar as for the continuous outcome, the scenario without association between mediator M and covariates X is obtained, with $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = (0,0,0,0,0,0,0,0)^T$ and variance $\sigma^2 = 1$. The scenario's with a moderate association between M and X are generated with $\beta_2 = (0.1,-0.2,0.8,0.15,0.2,-0.6,$ $0.25,-0.5)^T$. In the strong association setting, β_2 is changed to $(0.1,-0.2,1.15,$ $0.25,0.4,-1.3,0.5,-0.8)^T$. In terms of the strength of the association between outcome Y and covariates X, a setting with no association is created with $(\alpha_{00},\alpha_{01})$ equal to $(-0.8,0.8)$, $\alpha_{20} = (0,0,0,0,0,0,0,0)^T$ and $\alpha_{21} = (0,0,0,0,0,0,0,0)^T$. In the moderate scenario, α_{20} and α_{21} are modified to $(-0.1,0.1,0.2,-0.15,0,0,0,0)^T$ and $(0.1,-0.25,-0.2,0.15,0,0,0,0)^T$, and in the strong setting to $(-0.6,0.5,0.3,$ $-0.2,0,0,0,0)^T$ and $(0.4,-0.35,-0.3,0.7,0,0,0,0)^T$. To vary the strength of the association between outcome Y and mediator M, $(\alpha_{10},\alpha_{11})$ are adjusted from $(0,0)$

in the null scenario, to $(0.6, -0.8)$ in the moderate setting, and to $(0.8, -1)$ in the strong scenario.

2.A.5.2 Misspecification of the mediator model with a binary outcome

In the first mediator misspecification setting, where the mediator model is misspecified because the two squared terms were ignored, $(\alpha_{00}, \alpha_{01})$ equals $(-3, -2)$, $(\alpha_{10}, \alpha_{11})$ is set to $(1, 1.5)$, α_{20} to $(-0.1, 0, 0.1, -0.1, 0.05, 0, 0, 0)^T$ and α_{21} to $(-0.2, -0.2, 0, 0.05, 0.2, 0, 0, 0)^T$. The second scenario, characterized by an important predictor that is missing from the mediator model in the analysis, has parameter values $(\alpha_{00}, \alpha_{01}) = (-0.1, 0.05)$, $(\alpha_{10}, \alpha_{11}) = (-0.5, 0.5)$, $\alpha_{20} = (-0.2, 0.15, 0, 0, 0, 0, 0, 0)^T$ and $\alpha_{21} = (-0.15, 0.2, 0, 0, 0, 0, 0, 0)^T$. The binary outcome in the setting where the mediator distribution has outliers is generated using parameter values: $(\alpha_{00}, \alpha_{01}) = (-0.6, 0.6)$, $(\alpha_{10}, \alpha_{11}) = (0.3, -0.2)$, $\alpha_{20} = (-0.1, 0.1, 0.2, -0.15, 0, 0, 0, 0)^T$ and $\alpha_{21} = (0.1, -0.25, -0.2, 0.15, 0, 0, 0, 0)^T$. In the fourth and final scenario with a gamma mediator distribution $(\alpha_{00}, \alpha_{01})$ equals $(-2.5, -2)$, $(\alpha_{10}, \alpha_{11})$ equals $(-0.25, 0.25)$, α_{20} is set to $(0.15, -0.05, -0.1, 0, 0, 0, 0, 0)^T$ and α_{21} to $(0.05, 0.1, 0.05, 0, 0, 0, 0, 0)^T$.

2.A.5.3 Binary mediator

Covariates $X^* = (X_1, ..., X_8)^T$ are generated as follows: X_1, X_3 and $X_8 \sim \mathcal{N}(0, 1)$, X_4 and X_6 are Bernoulli with $P(X_4 = 1) = 0.3$ and $P(X_6 = 1) = 0.5$, and $X_2 = 0.2X_1 + 0.98U_1$, $X_5 = 0.1X_1 + 0.2X_3 + 0.97U_2$, and $X_7 = 0.1X_3 + 0.99U_3$, where $U_l \sim \mathcal{N}(0, 1)$, $l = 1, 2, 3$. The binary mediator M is generated as a Bernoulli variate obeying $\text{logit}\{P(M = 1|A, X)\} = \beta_0 + \beta_1 A + \beta_2^T X$ with X including the eight covariates X^* and one higher order term X_1^2. Parameter values are $\beta_0 = 0$, $\beta_1 = -1$, and $\beta_2 = (0.05, -0.1, 0.1, -0.2, 0, 0, 0, 0, 1)$. The mediator model used in the analyses is a logistic regression, including main effects of A and X^*. As a result, the mediator model is misspecified because a higher order term is ignored.

Continuous outcome The continuous outcome Y is drawn from a normal distribution with mean $E(Y|A, M, X) = \alpha_0 + \alpha_1 A + \alpha_2 M + \alpha_3^T X$, residual variance 4, X including the eight covariates of X^* and the following parameter values: $\alpha_0 = 1$,

$\alpha_1 = 2$, $\alpha_2 = -2$, $\alpha_3 = (0.8, 0.7, 0.55, -0.6, -0.25, 0, 0, 0)^T$. Table 2.7 summarizes the simulation results for the continuous outcome under mediator model misspecification for a binary mediator. As before, we observe a drastic efficiency benefit of the LE and RE estimators in comparison to the RMLE and TMLE estimators for the natural indirect effect estimate, even when the mediator model is misspecified. The MLE continues to be slightly more efficient, and not biased. As before, the larger efficiency of the MLE is due to a more restrictive outcome model, which does not allow for modification of the mediator and covariate effects by treatment. If the MLE does allow such effects to be included, the slight efficiency benefit it had disappears.

	Estimator	Bias	Emp SD	Mean SE	Coverage
Direct effect	RMLE	0.009	0.190	0.189	0.95
	LE	0.009	0.190	0.188	0.96
	RE	0.009	0.190	0.188	0.96
	MLE	0.009	0.184	0.183	0.95
	MLE_I	0.009	0.189	0.188	0.95
	TMLE_P	0.009	0.187	0.187	0.94
	TMLE_PP	0.009	0.187	0.169	0.91
	TMLE	0.010	0.187	0.169	0.91
Indirect effect	RMLE	0.006	0.160	0.160	0.96
	LE	-0.003	0.102	0.106	0.96
	RE	-0.003	0.102	0.106	0.96
	MLE	-0.003	0.094	0.096	0.96
	MLE_I	-0.003	0.102	0.103	0.95
	TMLE_P	0.006	0.155	0.157	0.96
	TMLE_PP	0.006	0.155	0.930	1.00
	TMLE	0.006	0.154	0.930	1.00

Table 2.7: Direct and indirect effect on a continuous outcome with binary mediator misspecification

Binary outcome The binary outcome Y is generated as Bernoulli according to $\text{logit}\{P(Y=1|A=a,M,X)=\alpha_{0a}+\alpha_{1a}M+\alpha_{2a}^{T}X\}$ with $a=0$ or 1. X included the eight covariates of X^*. The following parameter values are used: $(\alpha_{00},\alpha_{01})=(-2,-1)$, $(\alpha_{10},\alpha_{11})=(-0.5,1)$, $\alpha_{20}=(-0.1,0,-0.05,0.15,0.3,0,0,0)^T$ and $\alpha_{21}=(-0.2,-0.2,0.3,0.2,-0.1,0,0,0)^T$. As before, Table 2.8 shows that the MLE is biased when it does not take into account that the mediator and covariate effects on the outcome interact with treatment. Similarly, the 'non-parametric' TMLE, based on the flexible outcome model, again shows some bias. Also in terms of efficiency previous conclusions can be repeated.

	Estimator	Bias	Emp SD	Mean SE	Coverage
Direct effect	RMLE	0.000	0.041	0.040	0.95
	LE	-0.001	0.040	0.040	0.95
	RE	0.000	0.040	0.040	0.95
	MLE	-0.018	0.040	0.039	0.91
	MLE_I	-0.001	0.040	0.039	0.95
	TMLE_P	0.000	0.041	0.041	0.95
	TMLE_PP	0.000	0.041	0.040	0.94
	TMLE	-0.016	0.041	0.037	0.90
Indirect effect	RMLE	0.001	0.021	0.022	0.96
	LE	0.001	0.016	0.018	0.96
	RE	0.001	0.016	0.018	0.96
	MLE	0.019	0.013	0.013	0.70
	MLE_I	0.002	0.015	0.015	0.94
	TMLE_P	0.001	0.021	0.023	0.97
	TMLE_PP	0.001	0.021	0.043	1.00
	TMLE	0.015	0.018	0.041	1.00

Table 2.8: Direct and indirect effect on a binary outcome with binary mediator misspecification

2.A.5.4 Small sample

Now, we will evaluate the performance of the different proposed estimators through simulation analyses with 1000 runs for data sets of 50 observations. Covariates $X = (X_1, ..., X_4)^T$ are generated as follows: X_1 and X_3 are generated from a standard normal distribution, X_4 is Bernoulli with $P(X_4 = 1) = 0.5$ and $X_2 = 0.1X_1 + 0.2X_3 + 0.97U_1$, where $U_1 \sim \mathcal{N}(0, 1)$. The continuous mediator M is drawn from a normal distribution with residual variance 1 and mean $E(M|A, X) = \beta_0 + \beta_1 A + \beta_2^T X$ with $\beta_0 = 0$, $\beta_1 = 0.5$ and $\beta_2 = (0.1, -0.2, 0.15, -0.3)^T$.

	Estimator	Bias	Emp SD	Mean SE	Coverage
Direct effect	RMLE	0.002	0.683	0.650	0.94
	LE	0.007	0.660	0.626	0.94
	RE	0.008	0.649	0.623	0.94
	MLE	0.006	0.606	0.628	0.96
	MLE_I	0.012	0.664	0.776	0.97
	TMLE_P	-0.009	0.608	0.630	0.95
	TMLE_PP	-0.005	0.607	0.430	0.81
	TMLE	-0.342	0.857	0.440	0.65
Indirect effect	RMLE	0.013	0.538	0.555	0.96
	LE	0.002	0.409	0.426	0.95
	RE	0.001	0.407	0.426	0.95
	MLE	0.006	0.344	0.364	0.94
	MLE_I	0.002	0.404	0.477	0.96
	TMLE_P	0.014	0.451	0.52	0.98
	TMLE_PP	0.019	0.460	2.08	1.00
	TMLE	0.104	0.436	2.030	1.00

Table 2.9: Direct and indirect effect on a continuous outcome with binary mediator misspecification

Continuous outcome A continuous outcome Y is drawn from a normal distribution with mean $E(Y|A,M,X) = \alpha_0 + \alpha_1 A + \alpha_2 M + \alpha_3^T X$ and residual variance 4. Here, $\alpha_0 = 1$, $\alpha_1 = 2$, α_2 is set to -1 and α_3 to $(0.55, -0.7, 0.8, -0.55)^T$. Table 2.9 summarizes the simulation results for the continuous outcome. Relative to the RMLE and to a lesser extent the TMLE, the proposed efficient estimators (LE and RE) deliver drastic efficiency gains for the natural indirect effect, even with this small sample size. The relatively minor efficiency benefit of the MLE is again visible, but it disappears once the assumption that the mediator and covariate effects on the outcome do not interact with treatment is dropped. As in the data analysis, the 'non-parametric' TMLE shows its sensitivity to finite sample bias.

	Estimator	Bias	Emp SD	Mean SE	Coverage
Direct effect	RMLE	-0.004	0.168	0.155	0.91
	LE	-0.007	0.162	0.149	0.91
	RE	-0.005	0.163	0.149	0.91
	MLE	-0.018	0.159	0.151	0.93
	MLE_I	-0.006	0.164	0.160	0.97
	TMLE_P	-0.005	0.167	0.170	0.95
	TMLE_PP	-0.006	0.168	0.140	0.89
	TMLE	-0.040	0.146	0.110	0.87
Indirect effect	RMLE	0.002	0.090	0.101	0.97
	LE	0.003	0.067	0.080	0.98
	RE	0.003	0.068	0.081	0.98
	MLE	0.015	0.045	0.053	0.97
	MLE_I	0.003	0.065	0.078	0.99
	TMLE_P	0.001	0.091	0.140	1.00
	TMLE_PP	0.001	0.092	0.180	1.00
	TMLE	0.015	0.071	0.170	1.00

Table 2.10: Direct and indirect effect on a binary outcome with binary mediator misspecification

Binary outcome The binary outcome Y is generated as Bernoulli according to $\text{logit}\{P(Y=1|A=a,M,X)=\alpha_{0a}+\alpha_{1a}M+\alpha_{2a}^{T}X\}$ with $a=0$ or 1. The following parameter values are used: $(\alpha_{00},\alpha_{01})=(-0.3,0.1)$, $(\alpha_{10},\alpha_{11})=(0.1,-0.15)$, $\alpha_{20}=(-0.1,0.05,0.2,0)^{T}$ and $\alpha_{21}=(0.05,0,-0.1,0.1)^{T}$. Table 2.10 shows the simulation results for the binary outcome which are similar to those of the continuous outcome in terms of efficiency: large efficiency benefit for the LE and RE compared to the RMLE and the TMLE. The MLE is more efficient under the no treatment interaction assumption, but shows significant bias when this assumption does not hold. Here this is not that obvious, because the interactions were not really important predictors. When treatment interactions are included in the outcome model for the MLE, its efficiency gain disappears.

2.A.5.5 Outcome model misspecification

Finally, we will evaluate the performance of the different proposed estimators through simulation analyses with 1000 runs for data sets of 500 observations with outcome model misspecification. Covariates $X^{*}=(X_1,...,X_8)^{T}$ are generated as follows: X_1, X_3 and $X_8\sim\mathcal{N}(0,1)$, X_4 and X_6 are Bernoulli with $P(X_4=1)=0.3$ and $P(X_6=1)=0.5$, and $X_2=0.2X_1+0.98U_1$, $X_5=0.1X_1+0.2X_3+0.97U_2$, and $X_7=0.1X_3+0.99U_3$, where $U_l\sim\mathcal{N}(0,1)$, $l=1,2,3$. The continuous mediator M is drawn from a normal distribution with residual variance 1 and mean $E(M|A,X)=\beta_0+\beta_1A+\beta_2^{T}X$ with $\beta_0=0$, $\beta_1=1$ and $\beta_2=(0.1,-0.2,0.8,0.15,0.2,-0.6,0.25,-0.5)^{T}$.

Continuous outcome We will examine the effect of outcome misspecification via various simulation settings with misspecification of the outcome model. In the first and fifth setting, the continuous outcome Y is drawn from a normal distribution with mean $E(Y|A,M,X)=\alpha_0+\alpha_1A+\alpha_2M+\alpha_3^{T}X$, residual variance 4 and with X including the eight covariates X^{*} and one squared term X_8^2. Parameter value α_0 is set to 1, α_1 to 2, α_2 is set to -1 and $\alpha_3=(0.55,0.7,0.8,-0.55,-0.25,0,0,0,-0.75)^{T}$ in the first setting and to $\alpha_0=1$, $\alpha_1=2$, α_2 is set to -0.75 and $\alpha_3=(-0.5,0,0,0,0,0,0,0,1)^{T}$ in the fifth setting. As a result, the outcome models in the analyses are misspecified since the squared term was ignored.

Misspecification		RMLE	LE	RE	MLE	MLE_I	TMLE_PP	TMLE
Scenario 1	Bias	-0.002	-0.007	-0.007	-0.003	-0.007	-0.037	-0.030
	Emp SD	0.270	0.268	0.268	0.242	0.268	0.238	0.235
	Mean SE	0.250	0.248	0.248	0.225	0.250	0.187	0.225
	Coverage	0.94	0.93	0.93	0.93	0.94	0.85	0.92
Scenario 2	Bias	-0.003	-0.002	-0.003	-0.003	-0.002	-0.008	-0.002
	Emp SD	0.254	0.249	0.249	0.221	0.249	0.223	0.222
	Mean SE	0.238	0.235	0.235	0.210	0.237	0.164	0.202
	Coverage	0.94	0.94	0.94	0.93	0.94	0.82	0.91
Scenario 3	Bias	-0.003	-0.006	-0.006	-0.004	-0.006	-0.045	-0.086
	Emp SD	0.333	0.320	0.320	0.274	0.320	0.265	0.289
	Mean SE	0.300	0.289	0.289	0.251	0.293	0.190	0.202
	Coverage	0.93	0.93	0.93	0.92	0.93	0.81	0.84
Scenario 4	Bias	-0.003	-0.006	-0.006	-0.004	-0.006	-0.035	-0.044
	Emp SD	0.333	0.320	0.320	0.274	0.320	0.266	0.269
	Mean SE	0.300	0.289	0.289	0.251	0.293	0.181	0.195
	Coverage	0.93	0.93	0.93	0.92	0.93	0.80	0.89
Scenario 5	Bias	-0.008	-0.001	-0.001	-0.001	-0.001	0.026	0.010
	Emp SD	0.280	0.277	0.277	0.251	0.277	0.260	0.265
	Mean SE	0.270	0.268	0.268	0.243	0.269	0.201	0.250
	Coverage	0.94	0.94	0.94	0.95	0.94	0.85	0.93
Scenario 6	Bias	0.006	0.000	0.000	-0.001	-	-0.051	-0.113
	Emp SD	0.393	0.374	0.374	0.320	-	0.304	0.301
	Mean SE	0.364	0.346	0.346	0.295	-	0.220	0.221
	Coverage	0.95	0.93	0.94	0.93	-	0.82	0.82

Table 2.11: Direct effect on a continuous outcome with outcome misspecification

In the second, third, fourth and sixth scenario, the continuous outcome Y has mean $E(Y|A,M,X) = \alpha_0 + \alpha_1 A + \alpha_2 M + \alpha_3^T X + \alpha_4^T X \times M$ with residual variance 4 and X including the eight covariates X^*. In the second and third scenario, parameter values were set to $\alpha_0 = 1$, $\alpha_1 = 2$, α_2 is set to -1 and $\alpha_3 = (0.55, 0.7, 0.8, 0, -0.25, 0, 0, 0)^T$. Further in the second scenario, the parameter values of α_4 were set to $(0,0,0,1,0,0,0,0)^T$ and in the third scenario to $(0,0,0,0,0,0,0,1)^T$. In the fourth scenario, parameter values were set to $\alpha_0 = 1$, $\alpha_1 = 2$, α_2 is set to -0.75 and $\alpha_3 = (-0.5,0,0,0,0,0,0,0)^T$ and $\alpha_4 = (0,0,0,0,0,0,0,1)^T$. In the sixth setting, we left out the MLE with interaction terms for simplicity. Parameter values were set to $\alpha_0 = 1$, $\alpha_1 = 2$, α_2 is set to -1 and $\alpha_3 = (-0.5, 0.4, 0, -0.7, 0, 0, 0, 0)^T$ and $\alpha_4 = (0,0,0,0,-0.5,-0.2,-0.8,1)^T$. Outcome models are misspecified because they did not include the $X \times M$ interactions.

Table 2.12 shows that relative to the RMLE and to a lesser extent the TMLE's (i.e., in scenario's where covariates play an important role), the proposed efficient estimators (LE and RE) deliver drastic efficiency gains for the natural indirect effect, even with outcome misspecification. In terms of the natural direct effect (Table 2.11), the TMLE's seem to perform somewhat better in terms of efficiency in case of a continuous outcome. All estimators remain unbiased and confidence intervals reach the nominal level even with outcome model misspecification. With an R^2, indicating the importance of the forgotten squared or interaction term, ranging from 12% to 40%, this shows that the in case of continuous outcomes all competing estimators are fairly robust against outcome model misspecification.

Binary outcome In the first and fifth setting, the binary outcome Y is generated as Bernoulli according to logit$\{P(Y=1|A=a,M,X) = \alpha_{0a} + \alpha_{1a} M + \alpha_{2a}^T X\}$ with $a = 0$ or 1. X included the eight covariates of X^* and one squared term X_8^2. The following parameter values are used in the first scenario: $(\alpha_{00}, \alpha_{01}) = (-0.7, 0.1)$, $(\alpha_{10}, \alpha_{11}) = (0.2, -0.3)$, $\alpha_{20} = (-0.1, 0.1, -0.2, -0.15, 0, 0, 0, 0, 0.6)^T$ and $\alpha_{21} = (0.1, -0.15, -0.1, 0.15, 0, 0, 0, 0, 0.8)^T$. The fifth setting had parameter values $(\alpha_{00}, \alpha_{01}) = (-0.4, 0.1)$, $(\alpha_{10}, \alpha_{11}) = (-0.4, -0.6)$, $\alpha_{20} = (-0.1, 0, 0, 0, 0, 0, 0, 0, 0.5)^T$ and $\alpha_{21} = (0.1, 0, 0, 0, 0, 0, 0, 0, 1.5)^T$. As a result, the outcome models in the analyses are misspecified since the squared term was ignored. In the sec-

Misspecification		RMLE	LE	RE	MLE	MLE$_I$	TMLE$_{PP}$	TMLE
Scenario 1	Bias	0.007	0.008	0.008	0.005	0.008	0.032	0.029
	Emp SD	0.209	0.168	0.168	0.140	0.168	0.179	0.177
	Mean SE	0.215	0.172	0.172	0.135	0.170	1.118	1.117
	Coverage	0.96	0.95	0.95	0.95	0.95	1.00	1.00
Scenario 2	Bias	0.003	0.002	0.002	0.002	0.002	0.005	0.002
	Emp SD	0.192	0.154	0.154	0.119	0.155	0.155	0.161
	Mean SE	0.195	0.156	0.156	0.118	0.155	0.874	0.872
	Coverage	0.96	0.95	0.95	0.95	0.95	1.00	1.00
Scenario 3	Bias	0.003	0.008	0.008	0.006	0.008	0.036	0.044
	Emp SD	0.272	0.208	0.208	0.166	0.208	0.195	0.239
	Mean SE	0.279	0.206	0.206	0.158	0.204	1.444	1.438
	Coverage	0.96	0.94	0.94	0.94	0.93	1.00	1.00
Scenario 4	Bias	-0.002	0.007	0.007	0.005	0.007	0.024	0.011
	Emp SD	0.261	0.199	0.199	0.153	0.199	0.180	0.200
	Mean SE	0.258	0.197	0.197	0.147	0.195	1.389	1.385
	Coverage	0.96	0.95	0.94	0.94	0.94	1.00	1.00
Scenario 5	Bias	0.004	0.003	0.003	0.002	0.003	-0.030	-0.018
	Emp SD	0.204	0.173	0.173	0.131	0.173	0.171	0.178
	Mean SE	0.200	0.173	0.173	0.129	0.171	0.871	0.870
	Coverage	0.94	0.95	0.95	0.96	0.94	1.00	1.00
Scenario 6	Bias	-0.009	0.003	0.003	0.004	-	0.041	0.032
	Emp SD	0.350	0.253	0.253	0.201	-	0.239	0.288
	Mean SE	0.349	0.247	0.247	0.190	-	1.967	1.960
	Coverage	0.95	0.94	0.94	0.94	-	1.00	1.00

Table 2.12: Indirect effect on a continuous outcome with outcome misspecification

ond, third and fourth scenario, the binary outcome Y is generated as Bernoulli according to logit$\{P(Y=1|A=a,M,X)=\alpha_{0a}+\alpha_{1a}M+\alpha_{2a}^TX+\alpha_{3a}^TX\times M\}$ with $a=0$ or 1. X included the eight covariates of X^*. In the second scenario, parameter values were $(\alpha_{00},\alpha_{01})=(-0.4,0.1)$, $(\alpha_{10},\alpha_{11})=(0.4,-0.6)$, $\alpha_{20}=(-0.1,0.1,-0.2,0,0,0,0,0)^T$, $\alpha_{21}=(0.1,-0.15,-0.1,0,0,0,0,0)^T$, $\alpha_{30}=(0,0,0,0.5,0,0,0,0)^T$ and $\alpha_{31}=(0,0,0,1.5,0,0,0,0)^T$. In the third scenario, parameter values were $(\alpha_{00},\alpha_{01})=(-0.4,0.1)$, $(\alpha_{10},\alpha_{11})=(-0.4,-0.6)$, $\alpha_{20}=(-0.1,0.1,-0.2,0,0,0,0,0)^T$, $\alpha_{21}=(0.1,-0.15,-0.1,0,0,0,0,0)^T$, $\alpha_{30}=(0,0,0,0,0,0,0,0.5)^T$ and $\alpha_{31}=(0,0,0,0,0,0,0,1.5)^T$. In the fourth setting, parameter values were $(\alpha_{00},\alpha_{01})=(-0.4,0.1)$, $(\alpha_{10},\alpha_{11})=(-0.4,-0.6)$, $\alpha_{20}=(-0.1,0,0,0,0,0,0,0)^T$, $\alpha_{21}=(0.1,0,0,0,0,0,0,0)^T$, $\alpha_{30}=(0,0,0,0,0,0,0,0.5)^T$ and $\alpha_{31}=(0,0,0,0,0,0,0,1.5)^T$. In the sixth setting, the binary outcome Y is generated as Bernoulli according to logit$\{P(Y=1|A,M,X)=\alpha_0+\alpha_1A+\alpha_2M+\alpha_3^TX+\alpha_4^TX\times M\}$. X included the eight covariates of X^*. Parameter values were $\alpha_0=0$, $\alpha_1=0.6$, $\alpha_2=-1$, $\alpha_3=(0.1,0,0,0,0,0,0,0)^T$ and $\alpha_4=(0,0,0,0,0.7,0.3,1.5,0.6)^T$. Outcome models are misspecified because they did not include the $X\times M$ interactions.

Also for binary outcomes, we observe that relative to the RMLE and to a lesser extent the TMLE's, the proposed efficient estimators (LE and RE) perform better in terms of efficiency for the natural indirect effect, even with outcome misspecification (Table 2.14). Although there seems to be a bit more bias in scenario 3 and 4 for the RMLE, LE, RE, MLE's and 'partially parametric' TMLE, all confidence intervals approach the nominal level even with outcome model misspecification. We do believe that more severe outcome misspecification could possibly result in biased estimates and undercoverage.

Misspecification		RMLE	LE	RE	MLE	MLE_I	TMLE_PP	TMLE
Scenario 1	Bias	0.001	0.001	0.001	-0.050	0.001	0.001	-0.037
	Emp SD	0.053	0.053	0.053	0.051	0.053	0.053	0.050
	Mean SE	0.051	0.051	0.051	0.048	0.051	0.054	0.047
	Coverage	0.94	0.94	0.94	0.80	0.95	0.95	0.84
Scenario 2	Bias	0.002	0.002	0.001	-0.075	0.001	0.003	-0.024
	Emp SD	0.055	0.053	0.053	0.049	0.054	0.055	0.051
	Mean SE	0.056	0.055	0.055	0.050	0.055	0.061	0.052
	Coverage	0.95	0.95	0.95	0.67	0.96	0.96	0.92
Scenario 3	Bias	-0.015	-0.015	-0.015	0.001	-0.015	-0.018	-0.007
	Emp SD	0.055	0.054	0.054	0.048	0.054	0.058	0.049
	Mean SE	0.056	0.054	0.054	0.048	0.054	0.060	0.044
	Coverage	0.94	0.94	0.94	0.94	0.94	0.94	0.85
Scenario 4	Bias	-0.014	-0.014	-0.014	0.003	-0.014	-0.019	-0.007
	Emp SD	0.056	0.055	0.055	0.048	0.055	0.059	0.050
	Mean SE	0.056	0.055	0.055	0.048	0.055	0.060	0.043
	Coverage	0.94	0.94	0.94	0.95	0.94	0.94	0.84
Scenario 5	Bias	-0.001	0.000	0.000	-0.009	0.000	-0.001	-0.014
	Emp SD	0.047	0.047	0.047	0.044	0.047	0.048	0.044
	Mean SE	0.047	0.046	0.046	0.043	0.046	0.048	0.042
	Coverage	0.94	0.93	0.93	0.94	0.94	0.94	0.91
Scenario 6	Bias	0.001	0.002	0.002	-0.004	-	-0.004	-0.005
	Emp SD	0.050	0.049	0.049	0.045	-	0.045	0.044
	Mean SE	0.052	0.051	0.050	0.046	-	0.032	0.039
	Coverage	0.96	0.96	0.96	0.96	-	0.83	0.90

Table 2.13: Direct effect on a binary outcome with outcome misspecification

Misspecification		RMLE	LE	RE	MLE	MLE$_\mathrm{I}$	TMLE$_\mathrm{PP}$	TMLE
Scenario 1	Bias	0.000	0.000	0.000	0.051	0.000	0.000	0.030
	Emp SD	0.030	0.028	0.028	0.021	0.028	0.030	0.025
	Mean SE	0.031	0.029	0.029	0.021	0.029	0.068	0.067
	Coverage	0.96	0.96	0.96	0.32	0.95	1.00	1.00
Scenario 2	Bias	-0.001	-0.001	0.000	0.076	-0.001	-0.001	0.019
	Emp SD	0.035	0.031	0.031	0.023	0.031	0.035	0.031
	Mean SE	0.035	0.031	0.031	0.023	0.031	0.063	0.058
	Coverage	0.96	0.96	0.96	0.09	0.95	1.00	1.00
Scenario 3	Bias	0.015	0.015	0.015	-0.001	0.015	0.015	0.002
	Emp SD	0.037	0.032	0.032	0.022	0.032	0.037	0.030
	Mean SE	0.036	0.032	0.032	0.023	0.031	0.057	0.052
	Coverage	0.92	0.93	0.93	0.96	0.93	0.99	1.00
Scenario 4	Bias	0.015	0.015	0.015	-0.002	0.015	0.015	0.002
	Emp SD	0.036	0.032	0.032	0.022	0.032	0.036	0.028
	Mean SE	0.036	0.032	0.032	0.023	0.031	0.057	0.052
	Coverage	0.92	0.92	0.92	0.95	0.92	1.00	1.00
Scenario 5	Bias	0.002	0.002	0.002	0.011	0.002	0.001	0.009
	Emp SD	0.029	0.027	0.027	0.020	0.027	0.029	0.023
	Mean SE	0.029	0.027	0.027	0.019	0.026	0.070	0.069
	Coverage	0.95	0.95	0.95	0.91	0.94	1.00	1.00
Scenario 6	Bias	-0.002	-0.003	-0.002	0.003	-	0.002	0.000
	Emp SD	0.034	0.028	0.028	0.022	-	0.024	0.032
	Mean SE	0.034	0.029	0.029	0.021	-	0.066	0.065
	Coverage	0.96	0.95	0.95	0.94	-	1.00	1.00

Table 2.14: Indirect effect on a binary outcome with outcome misspecification

2.A.6 Other tables and figures

2.A.6.1 Simulation study: Correct model specification

Association		RMLE	LE	RE	MLE	MLE$_\text{I}$	TMLE$_\text{par}$	TMLE$_\text{PP}$	TMLE
$Y \sim M$: moderate	Bias	-0.005	-0.004	-0.004	-0.002	-0.004	-0.015	-0.015	-0.015
$Y \sim X$: moderate	Emp SD	0.235	0.234	0.234	0.211	0.234	0.212	0.212	0.213
$M \sim X$: null	Mean SE	0.221	0.219	0.219	0.200	0.221	0.202	0.169	0.172
	Coverage	0.93	0.93	0.93	0.93	0.93	0.93	0.86	0.86
$Y \sim M$: moderate	Bias	-0.005	-0.004	-0.004	-0.002	-0.004	-0.015	-0.015	-0.010
$Y \sim X$: moderate	Emp SD	0.235	0.234	0.234	0.211	0.234	0.212	0.211	0.210
$M \sim X$: strong	Mean SE	0.221	0.219	0.219	0.200	0.221	0.204	0.182	0.184
	Coverage	0.93	0.93	0.93	0.93	0.93	0.93	0.88	0.89
$Y \sim M$: null	Bias	-0.005	-0.004	-0.004	-0.002	-0.004	-0.005	-0.006	-0.005
$Y \sim X$: moderate	Emp SD	0.235	0.234	0.234	0.211	0.234	0.212	0.212	0.206
$M \sim X$: moderate	Mean SE	0.221	0.219	0.219	0.200	0.221	0.203	0.158	0.161
	Coverage	0.93	0.93	0.93	0.93	0.93	0.93	0.83	0.85
$Y \sim M$: strong	Bias	-0.005	-0.004	-0.004	-0.002	-0.004	-0.026	-0.026	-0.027
$Y \sim X$: moderate	Emp SD	0.235	0.234	0.234	0.211	0.234	0.212	0.212	0.213
$M \sim X$: moderate	Mean SE	0.221	0.219	0.219	0.200	0.221	0.204	0.194	0.197
	Coverage	0.93	0.93	0.93	0.93	0.93	0.93	0.91	0.90
$Y \sim M$: moderate	Bias	-0.005	-0.004	-0.004	-0.002	-0.004	-0.014	-0.014	-0.009
$Y \sim X$: null	Emp SD	0.235	0.234	0.234	0.211	0.234	0.210	0.210	0.199
$M \sim X$: moderate	Mean SE	0.221	0.219	0.219	0.200	0.221	0.202	0.167	0.164
	Coverage	0.93	0.93	0.93	0.93	0.93	0.93	0.85	0.87
$Y \sim M$: moderate	Bias	-0.005	-0.004	-0.004	-0.002	-0.004	-0.017	-0.017	-0.014
$Y \sim X$: strong	Emp SD	0.235	0.234	0.234	0.211	0.234	0.212	0.211	0.211
$M \sim X$: moderate	Mean SE	0.221	0.219	0.219	0.200	0.221	0.204	0.181	0.182
	Coverage	0.93	0.93	0.93	0.93	0.93	0.94	0.88	0.89

Table 2.15: Simulation results for direct effect on a continuous outcome under correct model specification.

Association		RMLE	LE	RE	MLE	MLE$_{\text{I}}$	TMLE$_{\text{P}}$	TMLE$_{\text{PP}}$	TMLE
$Y \sim M$: moderate	Bias	-0.001	-0.001	-0.001	-0.139	-0.002	-0.001	-0.001	-0.062
$Y \sim X$: moderate	Emp SD	0.049	0.049	0.049	0.052	0.050	0.050	0.050	0.052
$M \sim X$: null	Mean SE	0.049	0.049	0.049	0.051	0.050	0.050	0.051	0.047
	Coverage	0.95	0.94	0.94	0.21	0.95	0.94	0.95	0.71
$Y \sim M$: moderate	Bias	0.001	0.000	0.000	-0.103	0.000	0.001	0.001	-0.016
$Y \sim X$: moderate	Emp SD	0.050	0.046	0.047	0.050	0.047	0.051	0.051	0.048
$M \sim X$: strong	Mean SE	0.050	0.046	0.047	0.049	0.043	0.051	0.052	0.050
	Coverage	0.94	0.95	0.95	0.45	0.94	0.95	0.95	0.94
$Y \sim M$: null	Bias	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.002
$Y \sim X$: moderate	Emp SD	0.052	0.051	0.051	0.048	0.051	0.052	0.052	0.045
$M \sim X$: moderate	Mean SE	0.051	0.050	0.050	0.046	0.050	0.052	0.053	0.038
	Coverage	0.95	0.95	0.95	0.94	0.95	0.95	0.96	0.89
$Y \sim M$: strong	Bias	0.001	0.001	0.001	-0.142	0.000	0.001	0.001	-0.024
$Y \sim X$: moderate	Emp SD	0.052	0.049	0.049	0.053	0.051	0.053	0.053	0.051
$M \sim X$: moderate	Mean SE	0.051	0.048	0.048	0.051	0.047	0.052	0.052	0.050
	Coverage	0.94	0.94	0.94	0.21	0.93	0.94	0.94	0.93
$Y \sim M$: moderate	Bias	0.000	-0.001	-0.001	-0.124	-0.001	0.000	0.000	-0.032
$Y \sim X$: null	Emp SD	0.051	0.050	0.050	0.053	0.051	0.052	0.052	0.050
$M \sim X$: moderate	Mean SE	0.050	0.048	0.048	0.050	0.047	0.051	0.051	0.050
	Coverage	0.94	0.93	0.93	0.29	0.93	0.94	0.94	0.90
$Y \sim M$: moderate	Bias	-0.001	-0.001	-0.001	-0.113	-0.001	-0.001	-0.001	-0.026
$Y \sim X$: strong	Emp SD	0.052	0.048	0.049	0.051	0.050	0.052	0.052	0.050
$M \sim X$: moderate	Mean SE	0.050	0.047	0.047	0.049	0.045	0.051	0.051	0.049
	Coverage	0.93	0.94	0.94	0.37	0.93	0.94	0.94	0.92

Table 2.16: Simulation results for direct effect on a binary outcome under correct model specification.

2.A.6.2 Simulation study: Misspecification of the model for the mediator

Misspecification		RMLE	LE	RE	MLE	MLE$_{\mathrm{I}}$	TMLE$_{\mathrm{P}}$	TMLE$_{\mathrm{PP}}$	TMLE
Not included higher order terms	Bias	-0.004	-0.003	-0.003	-0.003	-0.004	0.020	0.021	0.012
	Emp SD	0.200	0.198	0.198	0.194	0.199	0.202	0.203	0.205
	Mean SE	0.190	0.188	0.188	0.183	0.189	0.194	0.187	0.186
	Coverage	0.94	0.94	0.94	0.92	0.93	0.93	0.92	0.91
Forgotten predictor	Bias	-0.001	-0.001	-0.001	-0.002	-0.001	0.092	0.091	0.014
	Emp SD	0.184	0.183	0.182	0.180	0.182	0.183	0.181	0.199
	Mean SE	0.184	0.183	0.183	0.180	0.182	0.186	0.171	0.167
	Coverage	0.95	0.95	0.95	0.95	0.94	0.93	0.90	0.89
Outliers in mediator distribution	Bias	0.011	0.011	0.011	0.011	0.011	0.015	0.016	0.002
	Emp SD	0.189	0.186	0.186	0.184	0.186	0.194	0.195	0.202
	Mean SE	0.186	0.185	0.185	0.181	0.185	0.191	0.184	0.183
	Coverage	0.95	0.95	0.95	0.95	0.95	0.95	0.91	0.90
Gamma mediator distribution	Bias	0.002	0.002	0.002	-0.002	0.002	-0.028	-0.028	-0.025
	Emp SD	0.204	0.203	0.203	0.197	0.203	0.203	0.204	0.202
	Mean SE	0.205	0.204	0.204	0.194	0.204	0.199	0.186	0.182
	Coverage	0.96	0.95	0.95	0.94	0.95	0.94	0.90	0.91

Table 2.17: Direct effect on a continuous outcome with mediator misspecification

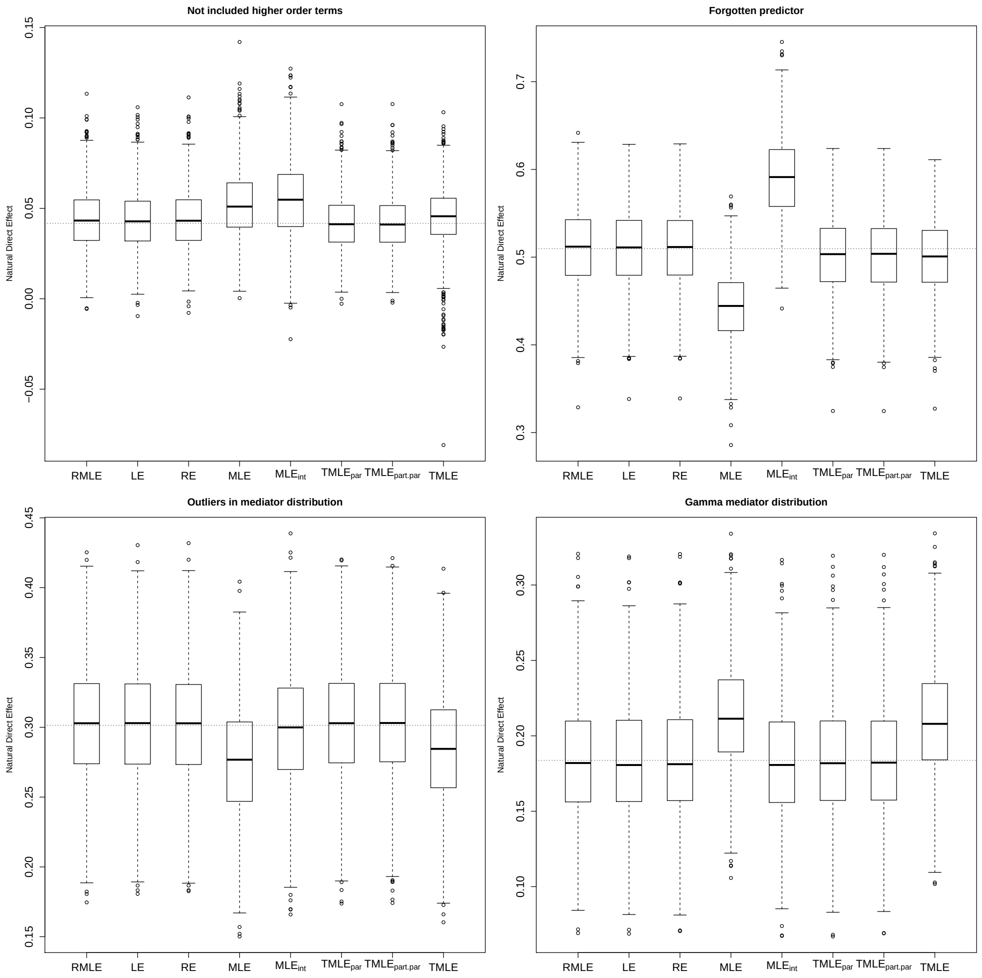


Figure 2.2: Direct effect on a binary outcome with mediator misspecification

Misspecification		RMLE	LE	RE	MLE	$\mathrm{MLE}_{\mathrm{I}}$	$\mathrm{TMLE}_{\mathrm{P}}$	$\mathrm{TMLE}_{\mathrm{PP}}$	TMLE
Not included higher order terms	Bias	0.000	0.000	0.000	0.009	0.011	-0.002	-0.002	0.002
	Emp SD	0.017	0.017	0.017	0.019	0.023	0.016	0.016	0.018
	Mean SE	0.017	0.017	0.017	0.019	0.025	0.017	0.016	0.015
	Coverage	0.95	0.95	0.95	0.92	0.98	0.95	0.96	0.91
Forgotten predictor	Bias	0.000	0.000	0.001	-0.066	0.081	-0.008	-0.008	-0.010
	Emp SD	0.047	0.046	0.046	0.041	0.047	0.045	0.045	0.044
	Mean SE	0.046	0.046	0.046	0.041	0.048	0.046	0.043	0.041
	Coverage	0.94	0.94	0.94	0.63	0.61	0.94	092	0.92
Outliers in mediator distribution	Bias	0.001	0.001	0.001	-0.026	-0.002	0.001	0.001	-0.018
	Emp SD	0.043	0.043	0.043	0.042	0.044	0.043	0.043	0.042
	Mean SE	0.045	0.044	0.044	0.044	0.045	0.046	0.045	0.041
	Coverage	0.95	0.95	0.95	0.92	0.96	0.96	0.95	0.91
Gamma mediator distribution	Bias	0.001	0.000	0.001	0.030	0.000	0.000	0.000	0.027
	Emp SD	0.040	0.039	0.039	0.036	0.039	0.040	0.040	0.038
	Mean SE	0.039	0.039	0.039	0.037	0.039	0.041	0.039	0.034
	Coverage	0.95	0.96	0.96	0.85	0.96	0.96	0.95	0.83

Table 2.18: Direct effect on a binary outcome with mediator misspecification

2.A.6.3 Data analysis

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	3.23	0.33	9.74	0.00
Age	-0.04	0.02	-1.92	0.07
Age^2	0.003	0.001	1.89	0.07
Evaluation	0.04	0.01	3.71	0.00

Table 2.19: Results of outcome model $E(Y|A=1,M,X)$

		Estimate	Std. Error	t value	Pr(>|t|)
Intercept		5.54	0.64	8.73	0.00
Rel. Status	Single				
	Relationship	-0.70	0.46	-1.54	0.14
Profession	Unemployed				
	Student	-0.97	0.63	-1.53	0.14
	Employed	-1.39	0.51	-2.73	0.01
Evaluation		0.01	0.01	0.59	0.56

Table 2.20: Results of outcome model $E(Y|A=0,M,X)$

		Estimate	Std. Error	t value	Pr(>|t|)
Intercept		4.81	0.49	9.89	0.00
Rel. Status	Single				
	Relationship	-0.83	0.44	-1.89	0.07
Education	Secondary				
	Higher	-1.16	0.44	-2.63	0.01
Gender	Male				
	Female	0.96	0.47	2.06	0.05

Table 2.21: Results of outcome model $E(Y|A=1,X)$

		Estimate	Std. Error	t value	Pr(>|t|)
Intercept		5.76	0.51	11.25	0.00
Profession	Unemployed				
	Student	-1.00	0.62	-1.61	0.12
	Employed	-1.40	0.50	-2.80	0.01
Rel. Status	Single				
	Relationship	-0.73	0.45	-1.62	0.12

Table 2.22: Results of outcome model $E(Y|A=0,X)$

		Estimate	Std. Error	t value	Pr(>|t|)
Intercept		4.85	0.34	14.36	0.00
Profession	Unemployed				
	Student	1.02	0.55	1.84	0.08
	Employed	-0.25	0.40	-0.62	0.54
Age		0.03	0.01	1.80	0.09

Table 2.23: Results of outcome model $E\left\{m(M,X;\eta^{*})|A=0,X\right\}$

CHAPTER 3

Efficient mediation analyses of time-to-event outcomes

This chapter is based on the following paper: Vandenberghe, S., Duchateau, L., Slaets, L., Bogaerts, J., and Vansteelandt, S. (2017). Surrogate marker analysis in cancer clinical trials through time-to-event mediation techniques. *Statistical Methods in Medical Research,* in press.

The meta-analytic approach is the gold standard for validation of surrogate markers, but has the drawback of requiring data from several trials. We refine modern mediation analysis techniques for time-to-event endpoints and apply them to investigate whether pathological complete response (pCR) can be used as a surrogate marker for disease-free survival (DFS) in the EORTC 10994/BIG 1-00 randomised phase 3 trial in which locally advanced breast cancer patients were randomised to either taxane or anthracycline based neoadjuvant chemotherapy. In the mediation analysis, the treatment effect is decomposed into an indirect effect via pCR and the remaining direct effect. It shows that only 4.2% of the treatment effect on disease-free survival after 5 years is mediated by the treatment effect on pCR. There is thus no evidence from our analysis that pCR is a valuable surrogate marker

to evaluate the effect of taxane versus anthracycline based chemotherapies on progression free survival of locally advanced breast cancer patients. The proposed analysis strategy is broadly applicable to mediation analyses of time-to-event endpoints, is easy to apply and outperforms existing strategies in terms of precision as well as robustness against model misspecification.

3.1 Introduction

This article is motivated by a secondary analysis of the EORTC 10994/BIG 1-00 randomised phase 3 trial in 1856 breast cancer patients. The study was originally set up to examine the interaction between P53 status of the patient and the treatment effect: the alternative hypothesis states that the hazard ratio of the taxane based regiment versus the anthracycline regimen is larger in the P53 positive patients as compared to the P53 negative patients (Bonnefoi et al. 2011). At the time of surgery, and thus after neoadjuvant chemotherapy, pathological complete response (pCR) status (i.e. a complete disappearance of any invasive cancer in the primary tumour and lymph nodes with the exception of very few scattered tumour cells) was determined. Our aim is to assess if pCR is an appropriate surrogate marker for the treatment effect on long-term clinically relevant outcomes, such as disease-free survival (DFS) and overal survival (OS), as previously suggested by Liedtke et al. (2008) and Mieog et al. (2007).

The validation of surrogate markers has been a longstanding popular research topic in the analysis of randomised trials (Daniels and Hughes 1997; Gail et al. 2000). Surrogate outcomes enable one to gain knowledge about the effect of treatment on the clinically relevant outcome from the effect of treatment on the surrogate, which is especially of interest in trials with low incidence rates for the clinically relevant outcomes. Since surrogate markers ideally provide early evidence about the effect of treatment, shorter follow-up periods and smaller sample sizes are required, which makes the resulting trials more cost-efficient. Event-free survival (Michiels et al. 2009) and disease-free survival (Sargent et al. 2005; Oba et al. 2013), for instance, have been examined as possible surrogate markers for overall survival in locally advanced head and neck cancer and resectable colon and gastric cancer respectively.

The meta-analytic approach is the gold standard approach for the validation of surrogate markers (Daniels and Hughes 1997; Buyse et al. 2000a; Alonso et al. 2015). It has the advantage of examining what surrogacy actually is about: does knowledge about the direction and strength of the effect of treatment on the surrogate enable one to predict the direction and strength of the effect of the treatment on the outcome. One large disadvantage is that it requires a fairly large number of studies before the results can be deemed reliable (Gail et al. 2000). A recent meta-analysis for example by Cortazar et al. (2014) studied whether pCR after a neoadjuvant chemotherapy regimen is an appropriate surrogate outcome for event-free survival (EFS) and overall survival (OS). Their pooled analysis was based on data from 10 international trials and showed no evidence of surrogacy. However, the results of a single trial in this meta-analysis containing an anti-HER2 treatment alongside chemotherapy, shifted the belief from an overall surrogacy to surrogacy restricted to the setting of targeted therapy for HER2-positive breast tumors. In 2012, the FDA issued a draft guidance about the use of pCR as an endpoint to support accelerated approval of a drug for high-risk, early stage breast cancer, such that patients can be provided (conditional) access to promising drugs while confirmatory clinical trials are being conducted (FDA 2014). In 2013, following this guideline, such accelerated approval was given to the anti-HER2 treatment Perjeta (FDA 2013), awaiting the phase III results. However, recently the ALTTO trial, focusing also on HER2 positive tumors, could not confirm its phase II pCR results and failed on its primary endpoint of disease-free survival (Goss et al. 2013). Therefore it is questionable whether pCR will remain a convincing red surrogate to the scientific community, since a surrogacy meta-analysis for anti-HER2 therapy is still lacking and moreover the clinical relevance of this endpoint for the patient is limited.

In light of these problems, we will use modern mediation analysis methods to examine the validity of pCR as a surrogate marker for disease-free survival (DFS). The advantage of this methodology is that only one trial is needed, but the disadvantage that it attempts to answer a more ambitious question: whether all of the treatment effect is mediated by the surrogate. The mediation analysis approach generalizes the proportion-explained approach for the validation of surrogate endpoints, better known as the Freedman approach (Freedman et al. 1992). This approach attempts to provide insight by contrasting the magnitude of the treatment effect with or without adjustment for the surrogate (see Tein and MacKinnon (2003) for such analysis in the context of time-to-event outcomes). Modern mediation analysis approaches accommodate one key problem of the Freedman approach: that adjusting for a marker in non-linear models (e.g. Cox regression models) may change the magnitude of the treatment effect even when the marker is not affected by the treatment (and thus a poor surrogate). This problem is known as non-collapsibility of treatment effects (Greenland et al. 1999; Martinussen and Vansteelandt 2013).

In this chapter, extending earlier work on continuous and binary outcomes (Vandenberghe et al. 2017a), we will introduce such modern mediation analysis approaches for the analysis of time-to-event endpoints and develop a novel approach that is simple, robust against model misspecification, and makes efficient use of the information in the data. We will study the usefulness of the proposed approach for the validation of surrogate markers, recognizing that the approach will be more broadly useful for mediation analysis of time-to-event endpoints. Indeed, such analyses are increasingly used and recommended in the analysis of randomised clinical trials to study why and how the treatment might achieve its effect on the outcome (Kraemer et al. 2002; Oakley et al. 2006). Understanding the different causal pathways through which the treatment influences the outcome may facilitate the development of innovative, better and more cost-effective treatments (Kraemer et al. 2002). For instance, Rochon et al. (2014) used a mediation analysis to examine the finding that patients treated in research-active institutions have better outcomes than patients treated in research-inactive institutions. Their study revealed that research-active institutions have superior patient survival due to a direct effect of research activity on survival, but also due to an indirect effect, since research-active institutions tended to make better use of surgery and chemotherapy. In the randomised trial of Pirl et al. (2012), early palliative care in patients with metastatic non-small-cell lung cancer was found to improve survival, but no sufficient evidence was found that this survival benefit was due to a reduction in depression scores.

The remainder of this chapter is organized as follows. In the next section, we give a detailed description of the EORTC 10994/BIG 1-00 randomised phase 3 trial dataset. In Section 3.3, we review existing mediation analysis approaches. In Section 3.4, these existing approaches are improved to increase the power provided that the models are correctly specified. Results of the method applied to the EORTC 10994/BIG 1-00 randomised phase 3 trial data will be presented in Section 3.5. We conclude with final remarks and ideas for future research in Section 3.6.

3.2 EORTC 10994/BIG 1-00 randomised phase 3 trial

The EORTC 10994/BIG 1-00 randomised phase 3 trial (Bonnefoi et al. 2011) examined whether patients with locally advanced breast cancer and a mutated P53 status were more sensitive to a taxane-based chemotherapy instead of a standard anthracycline regimen than women with wild-type P53. A secondary analysis of the trial data investigated the relationship between pCR and survival within intrinsic breast cancer subtypes (Bonnefoi et al. 2014). The trial included women with large operable or locally advanced/inflammatory breast cancer, for whom data were first collected on baseline covariates X_i (e.g., clinical nodal status, breast cancer subtype, $\ldots$). Subsequently, women were randomised (1:1) between two neoadjuvant treatment arms before undergoing primary surgery. They either received a standard anthracycline regimen ($A_i = 0$) or an experimental taxane-based regimen ($A_i = 1$), both over the course of approximately 15 weeks. At the time of surgery, pCR (M_i) was assessed by local pathologists. This is a binary variable coded as 1 if there is a complete disappearance of any invasive cancer in the primary tumour and lymph nodes with the exception of very few scattered tumour cells and 0 otherwise. The primary endpoint in our study is disease-free survival (DFS), defined as time from surgery to loco-regional recurrence, distant metastasis, death from any cause, or invasive contralateral breast cancer, whichever comes first. If none of these events occurred, women were censored at their last follow-up date. Median follow-up was 57 months.

A subgroup of the initial population of 1856 breast cancer patients was selected for this study based on several criteria (see the Appendix). After selection, 1546 (83%) patients appeared to be eligible for a mediation analysis with pCR. There are 766 eligible patients who were randomised to the standard treatment and 780 to the taxane-based regimen. For 459 of the 1546 eligible women (30%) an event was reported and 1087 women were thus censored at their last follow-up date. The analysis was ultimately limited to the 882 patients for whom complete data were available. Information is lost due to missingness in baseline covariates such as local PgR status (9.6%), p53 status (20.2%), histological grade (13.5%) and intrinsic breast cancer subtype (21.7%). Reassuringly, the intention-to-treat analysis on the entire subset (hazard ratio 0.77, 95% CI 0.64 to 0.93) versus the 882 patients with complete data (hazard ratio 0.75, 95% CI 0.59 to 0.96) showed similar hazard ratios for the effect of chemotherapy on DFS.

3.3 Available approaches

Like principal stratification methods for the validation of surrogate endpoints, modern developments to mediation analysis methods make use of potential outcome or counterfactual notation. We assume that counterfactual variables $M_i(a)$ and $T_i(a)$ exist for each patient $i = 1,...,n$ and each treatment group $a = 0,1$. Here, $M_i(0)$ and $T_i(0)$ correspond to the pCR status and duration of DFS that would have been observed for patient i, had she been randomised to the standard anthracycline regimen. As such, $M_i(0)$ and $T_i(0)$ are observed for control patients, but remain unobserved for patients on the experimental taxane-based chemotherapy arm. Furthermore, we use $T_i\{1,M_i(0)\}$ to represent the unobservable duration of DFS for patient i, had she been randomised to the experimental arm, but with the pCR status she would have had under control conditions. Using this potential outcome notation, we can define the direct effect of treatment on outcome on the risk ratio scale as

$$RR_d(t) = \frac{P[T_i\{1,M_i(0)\} > t]}{P[T_i\{0,M_i(0)\} > t]}, \tag{3.1}$$

which can be evaluated for each time $t > 0$. This is called a *natural direct effect* (Robins and Greenland 1992; Pearl 2001; VanderWeele and Vansteelandt 2009) since it captures the effect of treatment on outcome, while holding pCR at a value $M_i(0)$, which is the pCR value that would have been observed naturally for patient i if she would have been assigned to the standard anthracycline regimen. This natural direct effect shows what would happen to the duration of DFS if one could deliver the treatment without inducing a change in pCR status. A *natural indirect effect* (Robins and Greenland 1992; Pearl 2001; VanderWeele and Vansteelandt 2009) is correspondingly defined as

$$RR_m(t) = \frac{P[T_i\{1, M_i(1)\} > t]}{P[T_i\{1, M_i(0)\} > t]}, \tag{3.2}$$

for each time $t > 0$. It represents what would happen to the duration of DFS if pCR status were changed for each patient to the extent that it is affected by treatment for that patient. A marker with a natural direct effect of 1 at each time $t > 0$ and an indirect effect different from 1 at certain times $t > 0$, is likely a good surrogate marker, since all of the effect of treatment on the outcome then goes through the surrogate. Note that the product of the natural direct and indirect effects equals the total treatment effect $RR_{tot}(t) = \frac{P[T_i\{1, M_i(1)\} > t]}{P[T_i\{0, M_i(0)\} > t]}$.

It is important to realize that randomisation in itself is not enough to be able to disentangle the total effect of treatment into a natural direct and indirect effect. Even though patients were randomly assigned to the treatment, the mediator pCR is not randomly assigned. Direct and indirect effects conceptualize interventions on both randomised exposure and non-randomised mediator, and thus in particular demand control for confounding of the mediator-outcome association. Therefore, we will assume from now on that a set of baseline covariates X is sufficient to control for confounding of the association between pCR and DFS. This enables the use of the so-called mediation formula for the calculation of natural direct and indirect effects Pearl (2001). In practice, some of the confounders of the association between mediator and outcome may appear during the course of the study and might thus be influenced by the treatment themselves. The proposal in this article cannot handle these so-called intermediate confounders. In the discussion, we argue that the need for confounding adjustment is less essential when the only purpose is to evaluate whether M_i is a valuable surrogate of the effect of A_i on T_i.

Table 3.1: A restricted set of patients with baseline data, duration of DFS and model predictions.

Patient	1	2	3	4	5	6	7	8	9	10
Treatment A_i	0	1	1	0	0	0	1	0	1	0
pCR M_i	0	0	0	1	1	1	1	0	1	0
Menopausal status X_i	1	0	1	1	0	1	1	1	0	1
Event E_i	1	1	0	1	1	1	0	1	1	0
Duration of DFS T_i (days)	893	1162	937	108	1603	4476	10897	179	7342	122
$\hat{P}\{T_i(1) > 365 \mid X_i, M_i\}$	0.82	.	.	0.88	0.88	0.88	.	0.82	.	0.82
$\hat{P}\{T_i(1) > 730 \mid X_i, M_i\}$	0.66	.	.	0.76	0.78	0.76	.	0.66	.	0.66
$\hat{P}[T_i\{1, M_i(0)\} > 365 \mid X_i]$	0.83	0.85	0.83	0.83	0.85	0.83	0.83	0.83	0.85	0.83
$\hat{P}[T_i\{1, M_i(0)\} > 730 \mid X_i]$	0.69	0.72	0.69	0.69	0.72	0.69	0.69	0.69	0.72	0.69

The mediation formula (Pearl 2001, 2012) enables one to combine, in principle, arbitrary models for the mediator and outcome in order to obtain estimates of natural direct and indirect effects. For instance, Lange and Hansen (2011) show how to combine a normal regression model for a continuous mediator and an additive hazards model for the event time; in the case of rare events, VanderWeele (2011) makes progress using Cox proportional hazards models. Imai et al. (2010) relax the rare event assumption, but demand parametric survival models and represent natural direct and indirect effects in terms of mean differences in survival time, which is less ideal in the presence of skewness and censoring. Huang and Cai (2016) achieve greater flexibility by relying on semiparametric probit models for the event-time which combine well with linear models for the mediator. Considering the dominant concern for bias due to model misspecification in analyses of randomised studies, we will consider and develop more robust mediation analysis approaches in this article. In particular, like the proposal of Tchetgen Tchetgen (2011), we will merely demand correct specification of a model for the event time; however, to increase power relative to that proposal, we will additionally make use of a working model for the mediator.

To be concrete, let us first consider the data structure in Table 3.1, which shows artificial data for 10 breast cancer patients. As before, A_i is the randomised treatment indicator, X_i is a single binary baseline covariate (in practice multiple baseline covariates can be included), for example menopausal status, M_i defines pCR status, E_i is an indicator for a DFS event (1 if DFS occurred, 0 otherwise) and T_i is duration of DFS in days.

To calculate the numerator $P[T_i\{1, M_i(0)\} > t]$ of (3.1), we may proceed as follows:

1. Fit a Cox regression model for DFS given treatment arm, pCR and menopausal status

$$\lambda(t) = \lambda_0(t)\exp(\eta_1 A_i + \eta_2 M_i + \eta_3 X_i), \tag{3.3}$$

   using conventional software, where $\lambda_0(t)$ is the unspecified baseline hazard. Conventional software for fitting this model accommodates non-informative censoring, given treatment A_i, mediator pCR M_i and baseline covariate menopausal status X_i. Suppose that this results in the following parameter estimates for the log hazard ratio: $\hat{\eta}_1 = -0.95$, $\hat{\eta}_2 = -0.44$, and $\hat{\eta}_3 = 0.05$. In R these estimates can be obtained as follows: `fit.y <- coxph(Surv(T, event) ∼ A + M + X, data = cbind(A,T,event,X,M))`.
2. Use this Cox regression model to calculate the survival curve for all subjects, had they, possibly contrary to the fact, received the experimental taxane based regimen (rows 6 and 7 in Table 3.1):

$$\hat{P}\{T_i(1) > t|X_i, M_i\} = \exp\left\{-\hat{\Lambda}_0(t)\exp(\hat{\eta}_1 + \hat{\eta}_2 M_i + \hat{\eta}_3 X_i)\right\}. \tag{3.4}$$

   For the first patient, for example, the probability that the duration of DFS lasts longer than time t if she would have got the experimental treatment is calculated as: $\hat{P}\{T_1(1) > t|X_1, M_1\} = \exp[-\hat{\Lambda}_0(t)\exp\{(-0.95 \times 1) + (-0.44 \times 0) + (0.05 \times 1)\}]$, with $\hat{\Lambda}_0(t)$ the estimated cumulative baseline hazard which equalled 0.61 after 1 year and 0.36 after 2 years. In R we may calculate this as: `pp <- predictProb(fit.y,Surv(T,event), x = cbind(A=1 ,T,event,X,M), times = c(365,730)`.
3. Average the duration of DFS probabilities $\hat{P}\{T_i(1) > t|X_i, M_i\}$ for each time t

across all subjects in the control arm:

$$RMLE = \hat{P}[T_i\{1, M_i(0)\} > t] = \frac{1}{n_0} \sum_{i=1}^{n_0} \hat{P}\{T_i(1) > t | X_i, M_i\}, \tag{3.5}$$

with n_0 the number of subjects in the control group. This provides an estimate of the numerator $P[T_i\{1, M_i(0)\} > t]$ of (3.1). In our example, the duration of DFS probabilities of patients 1, 4, 5, 6, 8 and 10 are averaged to obtain an estimate for $P[T_i\{1, M_i(0)\} > t]$ at 1 and 2 years. The averaging is restricted to subjects in the standard anthracycline regimen because their observed value M_i for pCR equals $M_i(0)$. This is not the case for patients on the experimental arm, which is why we did not present the duration of DFS probabilities $\hat{P}\{T_i(1) > t | X_i, M_i\}$ for them in Table 3.1. Only using control patients to calculate $P[T_i\{1, M_i(0)\} > t]$ gives us a valid estimate, since both treatment groups are comparable due to randomization. In this case, the probability that the duration of DFS is larger than time t when patients would have got the experimental treatment but without influencing pCR is 0.85 and 0.71 after 1 and 2 years respectively. In R, we may calculate this as follows:

```
RMLE <- mean(cbind(pp,A)[A==0,1]).
```

We call estimator (3.5) the restricted maximum likelihood estimator (RMLE) because it uses predictions based on a model for the outcome in function of the treatment, mediator and confounders, but unlike the maximum likelihood approach of Lange and Hansen (2011) and VanderWeele (2011), no additional model for the mediator is used. This RMLE may be viewed as a special case of the more general IPW estimator in the paper of Tchetgen Tchetgen (2011) (pg. 5), which additionally allows control for measured confounders of the treatment-mediator and treatment-outcome association and can thus be used in observational studies.

3.4 Efficient estimator

A drawback of the estimation strategy in the previous section is that patients in the experimental arm do not contribute to the final calculation in (3.5), which makes the estimator inefficient. In view of this, we propose a locally efficient estimator

(LE) in this section that is guaranteed to be more efficient than the RMLE if all models used are correctly specified. To estimate the numerator $P[T_i\{1, M_i(0)\} > t]$ of (3.1), step 1 and 2 of the algorithm for the RMLE are repeated. Next, we proceed as follows:

3*. Regress the predictions of women in the standard anthracycline regimen from step 2 on baseline covariates X_i using a separate logistic regression in the control arm:

$$E[\hat{P}\{T_i(1) > t|X_i, M_i\}|A_i = 0, X_i] = \text{expit}(\alpha_0 + \alpha_1 X_i), \tag{3.6}$$

for each considered time t separately, with *expit* the inverse of the *logit* link function. This will later enable us to estimate the survival probabilities $\hat{P}[T_i\{1, M_i(0)\} > t|X_i]$ for subjects on the experimental arm, rather than just those on the control arm. Suppose that for time points 365 and 730 days, the logistic regression model (3.6) has parameter estimates $\hat{\alpha}_0 = 1.75$ and $\hat{\alpha}_1 = -0.13$, and $\hat{\alpha}_0 = 0.95$ and $\hat{\alpha}_1 = -0.14$ respectively. In R we may fit the logistic regression model as follows:

```
mod <- glm(pp ∼ X, data=cbind(A,T,event,X,M) [a==0,],
weights=1/(1-mean(A)), family=binomial).
```

4*. Use this logistic regression model at each time t to calculate $\hat{P}[T_i\{1, M_i(0)\} > t|X_i]$ for every subject (i.e. subjects from both the control and experimental arm) at each time t based on their observed baseline covariate values (Table 3.1):

$$\hat{P}[T_i\{1, M_i(0)\} > t|X_i] = \text{expit}(\hat{\alpha}_0 + \hat{\alpha}_1 X_i) \tag{3.7}$$

Because the model in step 3 is fitted only to women in the standard anthracycline regimen, its fitted values deliver estimates $\hat{P}[T_i\{1, M_i(0)\} > t|X_i]$ for 'control' levels of M_i. By randomization, the same model can be used to calculate $\hat{P}[T_i\{1, M_i(0)\} > t|X_i]$ for subjects on the experimental arm. Let us for example calculate the survival probability for the 10 patients in Table 3.1 for time points $t = 365$ and $t = 730$. For the first patient the probability that the duration of DFS lasts longer than $t = 365$ if she would have got the experimental treatment, but without influencing pCR, is calculated as $\hat{P}[T_1\{1,M_1(0)\} > 365|X_1] = \text{expit}(1.75 - 0.13 \times 1)$ and equals 0.83. The second patient, who was excluded from the calculation of the RMLE estimator, is now included in the analysis. The probability that the duration of DFS lasts longer than $t = 365$ if she would have got the experimental treatment, but without influencing pCR is calculated as $\hat{P}[T_2\{1,M_2(0)\} > 365|X_2] = \text{expit}(1.75 - 0.13 \times 0)$ and equals 0.72. In R this is:

```
pp.all <- predict(mod, newdata = cbind(A,T,event,X,
M,pp), type = "response").
```

5*. Average these probabilities $\hat{P}[T_i\{1,M_i(0)\} > t|X_i]$ for each time t across all subjects in both the control and experimental arm:

$$LE = \frac{1}{n}\sum_{i=1}^{n} \hat{P}[T_i\{1,M_i(0)\} > t|X_i]. \tag{3.8}$$

So now for time $t = 365$ and $t = 730$, we average these duration of DFS probabilities of all patients to get an estimate for $P[T_i\{1,M_i(0)\} > t]$. In this case, the probability that the duration of DFS lasts longer than time t when patients would get the experimental treatment but without influencing pCR, is 0.84 for $t = 365$ and 0.70 for $t = 730$, which is very similar to the results of the RMLE. In R we calculate this as:

```
LE <- mean(pp.all).
```

We show in the appendix that estimator (3.8) is consistent, in the sense that it converges to the true probability of duration of DFS as the sample size n increases, if the outcome model from step 1 is correctly specified. Quite surprisingly, this continues to be so even if the model from step 3* is misspecified. Additionally, this estimator is at least as efficient as the RMLE, when both models, the one from step 1 and step 3*, are correctly specified. We therefore called it locally efficient (LE) where ‘local’ emphasizes that efficiency is only guaranteed under correct model specification. That this estimator is more efficient than the RMLE can be intuitively expected because the estimator (3.8) averages across all subjects rather than only across the control subjects like (3.5). The estimation of the denominator of (3.1) and numerator of (3.2) follows by simply recoding the treatment accordingly in the previously described estimation steps. To report confidence intervals, we derived the analytic expression of the standard errors of the RMLE and LE estimator (Appendix) and confirmed that the LE is more efficient than the RMLE via simulations (see the Appendix). To facilitate the use of these estimators a user friendly R function is made available as part of the supplementary materials accompanying the paper, which includes calculation of the estimators, standard errors and confidence intervals.

3.5 Results of the EORTC 10994/BIG 1-00 trial

Before presenting results from our mediation analysis, we report the three relationships of interest between treatment, putative surrogate and clinical endpoint separately. pCR occurred in 283 of the 1546 (18%) women: 135 of the 766 (18%) in the standard anthracycline arm and 148 of the 780 (19%) in the experimental taxane based arm (odds ratio 1.09, 95% CI 0.85 to 1.42). The final multivariate Cox regression model, built using the general strategy in Collett (2003) to decide on the inclusion or exclusion of the available baseline characteristics age, height, weight, BMI, clinical nodal status, clinical tumour size, histological grade, histological type, local ER status, local PgR status, intrinsic breast cancer subtype, menopausal status and p53 status, is presented in Table 3.2. It shows a significant relationship between treatment and DFS (hazard ratio 0.65, 95% CI 0.51 to 0.84) and a significant interaction between pCR and clinical nodal status. The hazard ratio for DFS in the pCR 1 group for clinical nodal status N0 versus N1 is 0.37 (95% CI 0.16 to 0.85). The hazard ratio for DFS in the pCR 1 group for N0 versus N2-N3 is 0.21 (95% CI 0.02 to 1.80).

We used the RMLE and LE estimator to estimate the indirect effect of chemotherapy arm on DFS mediated via pCR and the remaining direct effect. The multivariate Cox regression model presented in Table 3.2 served as outcome model 3.3 for the first step of the RMLE and LE estimator. To obtain the LE estimator, we fit an additional logistic regression model 3.6 for each time *t* separately, which is shown for 2 of these time points in Table 3.3 and 3.4 of the Appendix. We used the same predictors as in the outcome model (Table 3.2), except those involving pCR and treatment.

Figure 3.1: Direct and indirect effect risk ratios of surviving the given time indicated on the X-axis with accompanying 95% point-wise confidence intervals

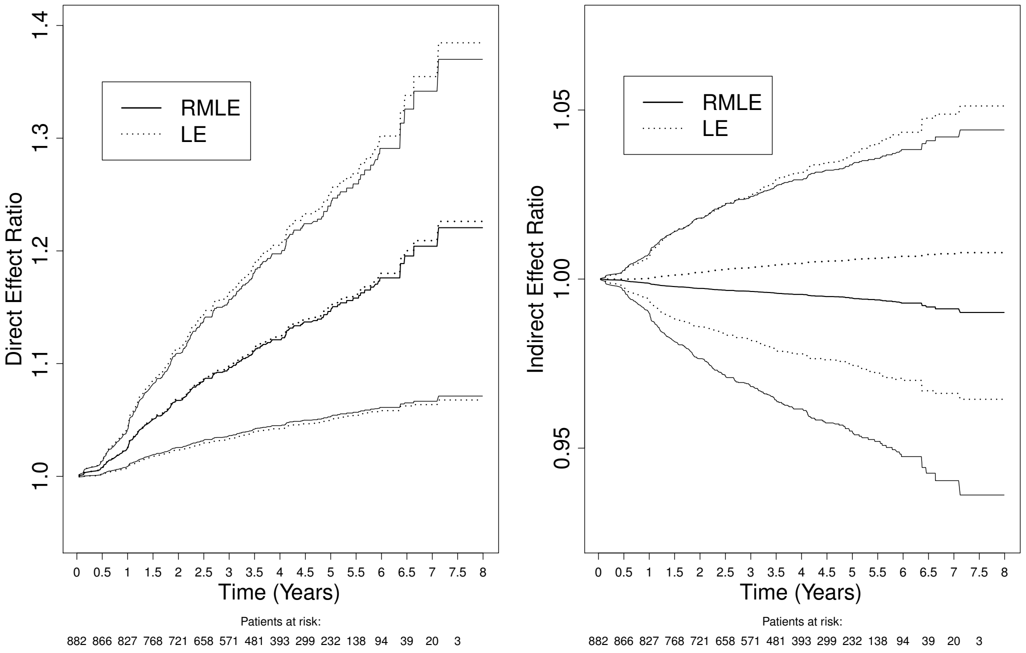


Figure 3.1 presents the direct effect ratio (3.1) of chemotherapy on DFS on the right-hand side and the indirect effect ratio (3.2) via pCR on the left-hand side. The RMLE estimator yields a direct effect of 1.148 (95% CI 1.053 to 1.242) and an indirect effect of 0.994 (95% CI 0.955 to 1.034) after 5 years. The LE estimator gave a similar direct effect of 1.151 (95% CI 1.050 to 1.251) and a more precise indirect effect of 1.006 (95% CI 0.975 to 1.037). In the Appendix, additional results of two sensitivity analyses are reported. First, an analysis on the complete set of eligible patients using the missing-indicator method (Greenland and Finkle 1995) on an outcome model that includes all important (possibly incomplete) covariates. Additionally, we evaluated if adjustment for a smaller number of baseline covariates would change our estimates. Thus, we also report results of a third analysis on the 882 patients with complete data that uses an outcome model without the incomplete covariates local PgR status, p53 status, histological grade and intrinsic breast cancer subtype. The overall results of our mediation analysis and the two sensitivity analyses (in the Appendix) are comparable. We thus conclude that the

Table 3.2: Results of the EORTC 10994/BIG 1-00 trial Cox regression model

		HR	95% CI	P-value
Treatment	Standard			
	Experimental	0.65	0.51 - 0.84	0.00
pCR	No			
	Yes	0.64	0.36 - 1.12	0.12
p53	Wild			
	Mutated	1.43	0.90 - 2.27	0.13
Local PgR status	Negative			
	Positive	0.86	0.53 - 1.40	0.54
Clin. nod. stat.				0.03[a]
	N0			
	N1	1.72	1.14 - 2.60	0.01
	N2 & N3	1.21	0.50 - 2.89	0.67
Clin. tum. size				0.00[a]
	T2			
	T3	2.83	0.93 - 8.60	0.07
	T4	7.65	1.52 - 38.48	0.01
Hist. grade				0.28[a]
	I			
	II	2.00	0.85 - 4.67	0.11
	III	2.92	1.08 - 7.92	0.04
BC subtype				0.00[a]
	Lum. A			
	Lum. B (HER2 neg.)	1.26	0.53 - 3.01	0.60
	Lum. B (HER2 pos.)	1.06	0.50 - 2.22	0.88
	HER2 pos. (non-lum.)	4.86	2.19 - 10.81	0.00
	Triple neg.	0.78	0.31 - 1.92	0.58
Clin. nod. stat. x pCR				0.03
	N0 x No			
	N1 x No			
	N2 & N3 x No			
	N0 x Yes			
	N1 x Yes	0.37	0.16 - 0.85	0.02
	N2 & N3 x Yes	0.21	0.02 - 1.80	0.15
Clin. tum. size x Hist. grade				0.04
	T2 x I			
	T3 x I			
	T4 x I			
	T2 x II			
	T3 x II	0.41	0.13 - 1.32	0.14
	T4 x II	0.28	0.05 - 1.58	0.15
	T2 x III			
	T3 x III	0.46	0.12 - 1.76	0.25
	T4 x III	0.09	0.01 - 0.65	0.02

[a] P-values based on a main effect model.

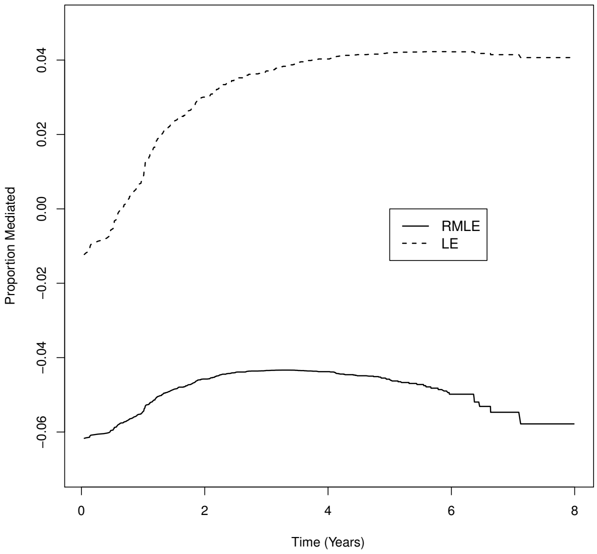
Table 3.2: Results of the EORTC 10994/BIG 1-00 trial Cox regression model

		HR	95% CI	P-value
p53 x BC subtype				0.05
	Wild x Lum. A			
	Mutated x Lum. A			
	Wild x Lum. B (HER2 neg.)			
	Mutated x Lum. B (HER2 neg.)	0.72	0.32 - 1.63	0.43
	Wild x Lum. B (HER2 pos.)			
	Mutated x Lum. B (HER2 pos.)	0.79	0.39 - 1.61	0.51
	Wild x HER2 pos. (non-lum.)			
	Mutated x HER2 pos. (non-lum.)	0.27	0.11 - 0.62	0.00
	Wild x Triple neg.			
	Mutated x Triple neg.	0.85	0.39 - 1.85	0.68
Clin. nod. stat. x Local PgR status				0.06
	N0 x Negative			
	N1 x Negative			
	N2 & N3 x Negative			
	N0 x Positive			
	N1 x Positive	0.86	0.50 - 1.49	0.60
	N2 & N3 x Positive	3.52	1.08 - 11.51	0.04
Clin. tum. size x BC subtype				0.00
	T2 x Lum. A			
	T3 x Lum. A			
	T4 x Lum. A			
	T2 x Lum. B (HER2 neg.)			
	T3 x Lum. B (HER2 neg.)	0.68	0.21 - 2.21	0.52
	T4 x Lum. B (HER2 neg.)	2.15	0.42 - 10.90	0.35
	T2 x Lum. B (HER2 pos.)			
	T3 x Lum. B (HER2 pos.)	3.06	1.29 - 7.24	0.01
	T4 x Lum. B (HER2 pos.)	2.53	0.83 - 7.76	0.10
	T2 x HER2 pos. (non-lum.)			
	T3 x HER2 pos. (non-lum.)	1.03	0.36 - 2.95	0.96
	T4 x HER2 pos. (non-lum.)	2.25	0.58 - 8.72	0.24
	T2 x Triple neg.			
	T3 x Triple neg.	3.12	1.22 - 7.97	0.02
	T4 x Triple neg.	14.81	4.36 - 50.35	0.00

[a] P-values based on a main effect model.

probability that the duration of DFS lasts longer than 5 years after administering the experimental taxane-based regimen is about 14.1% and 15.7% (11.7% and 11.5% in the analysis on the complete subset of eligible patients and 12.1% and 12% in the sensitivity analysis) larger for the RMLE and LE estimator respectively than when the anthracycline based regimen would be administered.

Figure 3.2: Proportion mediated for the given time indicated on the X-axis



The above analysis shows that a very small part of the total intention-to-treat effect is due to the effect via pCR. In particular, the mediation proportion, which is not a proportion in the real sense because it is not restricted to lie between 0 and 1, is calculated as follows (Ananth and VanderWeele 2011)

$$\frac{RR_{tot}(t) - RR_d(t)}{RR_{tot}(t) - 1} = \frac{P[T_i\{1, M_i(1)\} > t] - P[T_i\{1, M_i(0)\} > t]}{P[T_i\{1, M_i(1)\} > t] - P[T_i\{0, M_i(0)\} > t]}. \tag{3.9}$$

It shows that only 4.2% (0.3% in the analysis on the complete subset of eligible patients and 4.5% in the analysis using less covariates) of the treatment effect on the DFS risk difference is mediated by the treatment effect on pCR for the LE estimator after 5 years (Figure 3.2). In particular, changing the pCR status of a patient who got the taxane-based treatment to what it would have been in the anthracycline regimen decreases the probability that the duration of DFS lasts longer than 5 years with only 0.6% (based on the indirect effect of the LE estimator). That the indirect effect is small comes as no surprise, considering that no strong association was found between treatment and mediator pCR. We thus conclude, in line with the general results of Cortazar et al. (2014), that there is no evidence in favour of pCR being a surrogate for the effect of taxane versus anthracycline based neoadjuvant chemotherapy on disease-free survival.

From the confidence intervals, we observed no efficiency gain for the direct effect when using the LE estimator instead of the RMLE estimator, but a substantial 30% to 40% reduction in variance for the indirect effect (see also Figure 3.3 in the Appendix). This not surprising because the increased efficiency of our estimator is based on the model 3.6 for the mediator. Such a model enables more precise estimation of the exposure effect on the mediator, which is an essential component of the indirect effect and not the direct effect.

3.6 Discussion

We have proposed a novel, easy-to-apply estimator of the natural direct and indirect effect of a randomised treatment on a time-to-event outcome. Like the estimators of Tchetgen Tchetgen (2011), it quantifies the natural direct and indirect effect on the survival scale (instead of the hazard scale, as in Lange and Hansen (2011) and VanderWeele (2011)). This has the advantage of better interpretability (in view of the fact that hazard ratios lack causal interpretation, even in randomised trials (Hernan 2010; Aalen et al. 2015)), but the disadvantage of demanding a more high-dimensional graphical representation of the results. Our estimator improves the efficiency of a related estimator proposed by Tchetgen Tchetgen (2011), by extracting information from the potential outcomes to correct for baseline imbalances between randomised arms. Like the proposals by Lange and Hansen (2011) and VanderWeele (2011), this demands a model for (some transformation of) the mediator, but – unlike in those other proposals – misspecification of that model does not induce bias.

Our proposed methods are broadly applicable for mediation analysis of time-to-event endpoints in randomised trials. They were motivated by an application on the validation of surrogate markers, even though mediation is not a prerequisite for a good surrogate marker (Joffe and Greene 2009; VanderWeele and Vansteelandt 2013). In spite of this, in our opinion mediation analysis techniques are informative to study potential surrogate markers for the following reason. One can easily verify that the numerator of the mediation proportion (3.9) equals

$$\sum_m \sum_x P(T_i > t|M_i = m, A_i = 1, X_i = x)\{P(M_i = m|A_i = 1, X_i = x) - P(M_i = m|A_i = 0, X_i = x)\}P(X_i = x) \quad (3.10)$$

and the denominator equals the total treatment effect $P(T_i > t|A_i = 1) - P(T_i > t|A_i = 0)$. Regardless of whether the covariate set X includes all confounders of the mediator-outcome association, the calculated mediation proportion thus encodes a contrast of (some functional of) the treatment effect on the surrogate, $P(M_i = m|A_i = 1, X_i = x) - P(M_i = m|A_i = 0, X_i = x)$, and the treatment effect on the clinical endpoint, $P(T_i > t|A_i = 1) - P(T_i > t|A_i = 0)$, weighted by the extent to which the surrogate is predictive of the clinical endpoint. In particular, for a binary surrogate marker (coded 0 or 1) in the absence of covariates, it follows from (3.10) that the mediation proportion equals

$$p = \frac{\{P(T_i > t|M_i = 1, A_i = 1) - P(T_i > t|M_i = 0, A_i = 1)\}}{P(T_i > t|A_i = 1) - P(T_i > t|A_i = 0)} \times \{P(M_i = 1|A_i = 1) - P(M_i = 1|A_i = 0)\}. \quad (3.11)$$

Suppose now that this mediation proportion as well as the extent to which the surrogate is predictive of the clinical endpoint (as measured by $P(T_i > t|M_i = 1, A_i = 1) - P(T_i > t|M_i = 0, A_i = 1)$) are both stable across trials. Then the above expression suggests that a doubling of the treatment effect on the surrogate marker translates into a $p100\%$ relative increase in the treatment effect on the clinical endpoint. Although this interpretation holds regardless of whether or not covariate adjustment was considered, we do encourage adjustment for covariates sufficient to control for confounding of the association between surrogate marker and clinical endpoint, because evidence of an important mediation effect will generally add support to the validity of a candidate surrogate. More generally, note from (3.11) that when the mediation proportion equals zero, then it indicates that either there is no treatment effect on the surrogate (as is nearly the case in our study) or that the mediator and outcome are not associated (conditional on exposure and covariates); in both these cases, M_i would be poor as a surrogate endpoint. This approach based on mediation analysis thus has the advantage over the meta-analysis approach of needing only a single trial, but the disadvantage of being blind to between-study heterogeneity in treatment effects and thus less well generalisable.

When using our results to acquire insight into the treatment mechanism, the need to control for confounding of the mediator-outcome association becomes a key consideration. With concern for residual confounding by unmeasured variables, sensitivity analysis techniques (Tchetgen Tchetgen 2011) may be applied. In our analysis, a remaining concern is that some patients died or experienced the event before surgery took place and thus before the mediator was assessed. If this occurs, then the mediation analysis must be limited to the subgroup of patients who are alive at the time at which the mediator is assessed. In this subgroup, the two treatment arms may not longer be comparable. In future work, we will extend the proposed techniques to account for this.

3.A Appendix

3.A.1 General theory

Let $\hat{S}\{t|A_i=1,M_i,X_i;\hat{\Lambda}_0(t,\hat{\eta}),\hat{\eta}\}$ be a consistent estimator of $P(T>t|A_i=1,M_i,X_i)$ obtained under Cox regression model 3.3. Note that the following theory can be generalized to other Cox regression models though. Suppose that the randomisation probabilities $\pi_i \equiv P(A_i=1|X_i;\alpha)$ are known and that also the nuisance parameters η and $\Lambda_0(t,\eta)$ of the survival model are known. Let $\hat{\pi}_i$, $\hat{\eta}$ and $\hat{\Lambda}_0(t,\hat{\eta})$ be consistent estimators of π_i, η and $\Lambda_0(t,\eta)$ respectively. Further, $dN_j(s)$ indicates if an event is observed at time s for unit j. We will use $\hat{\Lambda}_0(t,\hat{\eta})$ to explicate the dependence on $\hat{\eta}$ of the cumulative baseline hazard under model 3.3. Now, consider the restricted maximum likelihood estimator (RMLE):

$$\begin{aligned}
\hat{S}_{1M_0}(t) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-A_i}{P(A_i=0|X_i;\hat{\alpha})}\, S\{t|A_i=1,M_i,X_i;\hat{\Lambda}_0(t,\hat{\eta}),\hat{\eta}\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-A_i}{1-\hat{\pi}_i}\exp\{-\hat{\Lambda}_0(t,\hat{\eta})\,\exp(\hat{\eta}_1+\hat{\eta}_2 M_i+\hat{\eta}_3 X_i)\}\right] \\
&\equiv \hat{S}\{t,\hat{\pi},\hat{\eta},\hat{\Lambda}_0(t,\hat{\eta})\},
\end{aligned}$$

where

$$\hat{\Lambda}_0(t,\eta) = \int_0^t \frac{\sum_{j=1}^{n} dN_j(s)}{\sum_{j=1}^{n} I(T_j \geq s)\exp(\eta_1 A_j+\eta_2 M_j+\eta_3 X_j)},$$

the Breslow estimate of the baseline cumulative hazard function. This restricted maximum likelihood estimator of $\theta = S_{1M_0}(t)$ can be thus be obtained by solving an estimating equation of the form

$$\begin{aligned}
0 &= U_i\{\theta,\alpha,\Lambda_0(t,\eta),\eta\} \\
&= \sum_{i=1}^{n}\frac{1-A_i}{1-P(A_i=1|X_i;\alpha)}\, S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}-\theta,
\end{aligned}$$

with $\alpha,\Lambda_0(t,\eta)$ and η substituted by consistent estimators.

To derive the asymptotic distribution of $\hat{\theta}$, we make the following decomposition for each time t:

$$
\begin{aligned}
\sqrt{n}\,\{\hat{S}_{1M_0}(t) - S_{1M_0}(t)\} &= \sqrt{n}\,[\hat{S}\{t,\hat{\pi},\hat{\eta},\hat{\Lambda}_0(t,\hat{\eta})\} - \hat{S}\{t,\pi,\hat{\eta},\hat{\Lambda}_0(t,\hat{\eta})\}] \quad (3.12)\\
&+\ \sqrt{n}\,[\hat{S}\{t,\pi,\hat{\eta},\hat{\Lambda}_0(t,\hat{\eta})\} - \hat{S}\{t,\pi,\eta,\hat{\Lambda}_0(t,\eta)\}] \quad (3.13)\\
&+\ \sqrt{n}\,[\hat{S}\{t,\pi,\eta,\hat{\Lambda}_0(t,\eta)\} - \hat{S}\{t,\pi,\eta,\Lambda_0(t,\eta)\}] \quad (3.14)\\
&+\ \sqrt{n}\,[\hat{S}\{t,\pi,\eta,\Lambda_0(t,\eta)\} - S_{1M_0}(t)] \quad (3.15)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{i1} + U_{i2} + U_{i3} + U_{i4}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i,
\end{aligned}
$$

where in each step, we account for a different part of the uncertainty in the RMLE caused by estimating a particular model parameter and $S_{1M_0}(t)$ equals

$$
E\left[\frac{1-A_i}{1-\pi_i}\exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\right].
$$

Consider the first term (3.12) and note by Taylor expansion that

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{i1} &\equiv \sqrt{n}[\hat{S}\{t,\hat{\pi},\hat{\eta},\hat{\Lambda}_0(t,\hat{\eta})\} - \hat{S}\{t,\pi,\hat{\eta},\hat{\Lambda}_0(t,\hat{\eta})\}]\\
&= \sqrt{n}\,\frac{\partial}{\partial \pi}\hat{S}\{t,\pi,\eta,\Lambda_0(t,\eta)\}\,(\hat{\pi}-\pi) + o_p(1)\\
&= E\left[\frac{1-A_i}{(1-\pi_i)^2}\exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\right]\times\\
&\quad \sqrt{n}(\hat{\pi}-\pi) + o_p(1)\\
&= E\left[\frac{1-A_i}{(1-\pi_i)^2}\exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\right]\times\\
&\quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(A_i-\pi) + o_p(1)
\end{aligned}
$$

Thus,

$$
U_{i1} = E\left[\frac{1-A_i}{(1-\pi_i)^2}\exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\right](A_i-\pi_i).
$$

Consider the next term (3.13) and note by Taylor expansion that

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{i2} &\equiv \sqrt{n}[\hat{S}\{t,\pi,\hat{\eta},\hat{\Lambda}_0(t,\hat{\eta})\} - \hat{S}\{t,\pi,\eta,\hat{\Lambda}_0(t,\eta)\}] \\
&= \sqrt{n}\,\frac{\partial}{\partial\eta}\hat{S}\{t,\pi,\eta,\Lambda_0(t,\eta)\}\;(\hat{\eta}-\eta) + o_p(1) \\
&= \sqrt{n}(\hat{\eta}-\eta)\times\Bigg(-E\left[\frac{1-A_i}{1-\pi_i}\exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\times\right. \\
&\qquad \left.\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\begin{pmatrix}1\\ M_i\\ X_i\end{pmatrix}\right] + E\left[\frac{1-A_i}{1-\pi_i}\times\right. \\
&\qquad \exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\times \\
&\qquad \left.\int_0^{\tau}\frac{\sum_{j=1}^{n} dN_j(s)\;\sum_{j=1}^{n} I(T_j\geq s)\;\exp(\eta_1 A_j+\eta_2 M_j+\eta_3 X_j)\begin{pmatrix}A_j\\ M_j\\ X_j\end{pmatrix}}{\{\sum_{j=1}^{n} I(T_j\geq s)\;\exp(\eta_1 A_j+\eta_2 M_j+\eta_3 X_j)\}^2}\right]\Bigg) \\
&\qquad + o_p(1)
\end{aligned}
$$

It follows from standard survival theory that

$$
\sqrt{n}\left(\begin{pmatrix}\hat{\eta}_1\\ \hat{\eta}_2\\ \hat{\eta}_3\end{pmatrix} - \begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left(\frac{\partial V_i(\eta)}{\partial\eta}\right)^{-1} V_i(\eta) + o_p(1),
$$

where $V_i(\eta)$ is the partial score, i.e.

$$
V_i(\eta) = \int_0^{\tau}\left[\begin{pmatrix}A_i\\ M_i\\ X_i\end{pmatrix} - \frac{E\left\{\begin{pmatrix}A_i\\ M_i\\ X_i\end{pmatrix}\;I(T_i\geq u)\;\exp(\eta_1 A_i+\eta_2 M_i+\eta_3 X_i)\right\}}{E\left\{I(T_i\geq u)\;\exp(\eta_1 A_i+\eta_2 M_i+\eta_3 X_i)\right\}}\right] dM_i(u)
$$

with $dM_i(u) = dN_i(u) - I(T_i \geq u)\ d\Lambda_0(u,\eta)\ \exp(\eta_1 A_i + \eta_2 M_i + \eta_3 X_i)$ the martingale and τ the end-of-study time. Further $-E\left(\frac{\partial V_i(\eta)}{\partial\eta}\right)$ equals the expected information matrix, which is the covariance matrix of the score residuals $V_i(\eta)$. Thus,

$$U_{i2} = E\left(\frac{\partial V_i(\eta)}{\partial\eta}\right)^{-1} V_i(\eta)\times\Bigg(-E\left[\frac{1-A_i}{1-\pi_i}\exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\times\right.$$

$$\left.\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\begin{pmatrix}1\\M_i\\X_i\end{pmatrix}\right] + E\left[\frac{1-A_i}{1-\pi_i}\times\right.$$

$$\exp\{-\Lambda_0(t,\eta)\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\times$$

$$\left.\int_0^\tau \frac{\sum_{j=1}^n dN_j(s)\ \sum_{j=1}^n I(T_j\geq s)\ \exp(\eta_1 A_j+\eta_2 M_j+\eta_3 X_j)\begin{pmatrix}A_j\\M_j\\X_j\end{pmatrix}}{\{\sum_{j=1}^n I(T_j\geq s)\ \exp(\eta_1 A_j+\eta_2 M_j+\eta_3 X_j)\}^2}\right]\Bigg)$$

$$+o_p(1).$$

Next, consider the third term (3.14) and note by Taylor expansion that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n U_{i3} &\equiv \sqrt{n}[\hat{S}\{t,\pi,\eta,\hat{\Lambda}_0(t,\eta)\} - \hat{S}\{t,\pi,\eta,\Lambda_0(t,\eta)\}]\\ &= \sqrt{n}\ \frac{\partial}{\partial\Lambda_0(t,\eta)}\ \hat{S}\{t,\pi,\eta,\Lambda_0(t,\eta)\}\ \{\hat{\Lambda}_0(t,\eta)-\Lambda_0(t,\eta)\} + o_p(1)\\ &= -E\left[\frac{1-A_i}{1-\pi_i}\ \exp\{-\Lambda_0(t,\eta)\ \exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\}\times\right.\\ &\qquad\left.\exp(\eta_1+\eta_2 M_i+\eta_3 X_i)\right]\sqrt{n}\ \{\hat{\Lambda}_0(t,\eta)-\Lambda_0(t,\eta)\}+o_p(1),\end{aligned}$$

where

$$\begin{aligned}&\sqrt{n}\ \{\hat{\Lambda}_0(t,\eta)-\Lambda_0(t,\eta)\}\\ &\quad=\sqrt{n}\left\{\int_0^\tau \frac{\sum_{j=1}^n dN_j(u)}{\sum_{j=1}^n I(T_j\geq u)\ \exp(\eta_1 A_j+\eta_2 M_j+\eta_3 X_j)} - \int_0^\tau d\Lambda_0(u,\eta)\right\}\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{n}\Bigg\{ \int_0^\tau \frac{\sum_{j=1}^n dN_j(u)}{\sum_{j=1}^n I(T_j \geq u)\ \exp(\eta_1 A_j + \eta_2 M_j + \eta_3 X_j)} - \\
&\qquad \int_0^\tau \frac{d\Lambda_0(u,\eta)\ \sum_{j=1}^n I(T_j \geq u)\ \exp(\eta_1 A_j + \eta_2 M_j + \eta_3 X_j)}{\sum_{j=1}^n I(T_j \geq u)\ \exp(\eta_1 A_j + \eta_2 M_j + \eta_3 X_j)} \Bigg\} \\
&= \sqrt{n}\Bigg\{ \int_0^\tau \frac{\sum_{j=1}^n dN_j(u) - d\Lambda_0(u,\eta)\ I(T_j \geq u)\ \exp(\eta_1 A_j + \eta_2 M_j + \eta_3 X_j)}{\sum_{j=1}^n I(T_j \geq u)\ \exp(\eta_1 A_j + \eta_2 M_j + \eta_3 X_j)} \Bigg\} \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^n \left[\int_0^\tau \frac{dM_j(u)}{E\{I(T_j \geq u)\ \exp(\eta_1 A_j + \eta_2 M_j + \eta_3 X_j)\}} \right] + o_p(1).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
U_{i3} = &-E\left[\frac{1-A_i}{1-\pi_i}\ \exp\{-\Lambda_0(t,\eta)\ \exp(\eta_1 + \eta_2 M_i + \eta_3 X_i)\} \times \right. \\
&\left. \exp(\eta_1 + \eta_2 M_i + \eta_3 X_i) \right] \int_0^\tau \frac{dM_j(u)}{E\{I(T_j \geq u)\ \exp(\eta_1 A_j + \eta_2 M_j + \eta_3 X_j)\}}.
\end{aligned}
$$

And finally, the last term (3.15) is readily obtained as

$$
\begin{aligned}
\frac{1}{\sqrt{n}} \sum_{i=1}^n U_{i4} &\equiv \sqrt{n}[\hat{S}\{t, \pi, \eta, \Lambda_0(t,\eta)\} - S_{1M_0}(t)] \\
&= \sqrt{n}\Bigg(\frac{1}{n} \sum_{i=1}^n \left[\frac{1-A_i}{1-\pi_i}\ \exp\{-\Lambda_0(t,\eta)\ \exp(\eta_1 + \eta_2 M_i + \eta_3 X_i)\} \right] \\
&\quad - S_{1M_0}(t) \Bigg).
\end{aligned}
$$

Thus,

$$
U_{i4} = \frac{1-A_i}{1-\pi_i}\ \exp\{-\Lambda_0(t,\eta)\ \exp(\eta_1 + \eta_2 M_i + \eta_3 X_i)\}.
$$

To obtain standard errors, we use that

$$
\begin{aligned}
\mathrm{Var}[\sqrt{n}\ \{\hat{S}_{1M_0}(t) - S_{1M_0}(t)\}] &= \mathrm{Var}(\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i) + o_p(1) \\
&= \mathrm{Var}(U_i) + o_p(1).
\end{aligned}
$$

Thus,

$$\begin{aligned}\mathrm{Var}\{\hat{S}_{1M_0}(t)\} &\approx \frac{\mathrm{Var}(U_i)}{n}\\ &= \frac{\mathrm{Var}(U_{i1}+U_{i2}+U_{i3}+U_{i4})}{n}.\end{aligned}$$

Now that we know the restricted maximum likelihood estimator and its standard errors, we can make it more efficient by solving the following estimation equation instead

$$\begin{aligned}0 =& \sum_{i=1}^{n} U_i\{\theta,\alpha,\Lambda_0(t,\eta),\eta\}\\ =& \sum_{i=1}^{n} \frac{1-A_i}{1-P(A_i=1|X_i;\alpha)} S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}-\theta\\ &+\ d(X_i)\left\{A_i-P(A_i=1|X_i;\alpha)\right\},\end{aligned}$$

for some index function $d(X_i)$.

Similar as before, the variance of the solution $\hat{\theta}$ to this equation equals 1 over n times the variance of $U_i\{\theta,\alpha,\Lambda_0(t,\eta),\eta\}$. Assuming that randomisation probabilities π_i and also the nuisance parameters η and $\Lambda_0(t,\eta)$ of the survival model are known, the optimal choices of index function $d(X_i)$ may be obtained by minimising the variance of $U_i\{\theta,\alpha,\Lambda_0(t,\eta),\eta\}$ w.r.t. $d(X_i)$. It may be obtained by population least squares projection of $(1-A_i)-S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}/\left\{1-P(A_i=1|X_i;\alpha)\right\}$ onto $A_i-P(A_i=1|X_i;\alpha)$. $d_{\mathrm{opt}}(X_i)$ thus equals

$$\begin{aligned}&=-\frac{E\left[\frac{1-A_i}{1-P(A_i=1|X_i;\alpha)}S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}\left\{A_i-P(A_i=1|X_i;\alpha)\right\}|X_i\right]}{E\left[\left\{A_i-P(A_i=1|X_i;\alpha)\right\}^2|X_i\right]}\\ &=\frac{E\left[(1-A_i)\frac{P(A_i=1|X_i;\alpha)}{1-P(A_i=1|X_i;\alpha)}S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|X_i\right]}{P(A_i=1|X_i;\alpha)\left\{1-P(A_i=1|X_i;\alpha)\right\}}\\ &=\frac{E\left[(1-A_i)S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|X_i\right]}{\left\{1-P(A_i=1|X_i;\alpha)\right\}^2}\\ &=\frac{E\left[S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|A_i=0,X_i\right]}{1-P(A_i=1|X_i;\alpha)}.\end{aligned}$$

Calculating the efficient estimator of $S_{1M_0}(t)$ thus requires a working model for the conditional expectation $E\left[S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|A_i=0,X_i\right]$, which we here more generally formalise as

$$E\left[S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|A_i=0,X_i\right]=z(X_i;\gamma),$$

where $z(X;\gamma)$ is a known function. For given estimator $\hat{\gamma}$ of γ, the efficient estimator is then obtained as

$$\begin{aligned}\hat{\theta}=&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1-P(A_i=1|X_i;\alpha)}\Bigg[(1-A_i)S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}\\&+z(X_i;\hat{\gamma})\{A_i-P(A_i=1|X_i;\alpha)\}\Bigg] \qquad (3.16)\\=&\frac{1}{n}\sum_{i=1}^{n}z(X_i;\hat{\gamma})+\frac{1-A_i}{1-P(A_i=1|X_i;\alpha)}\times\\&\Bigg[S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}-z(X_i;\hat{\gamma})\Bigg]. \qquad (3.17)\end{aligned}$$

Upon noting that the estimation equation solved by the optimal estimator can be written as 3.17, it is clear that $z(X_i;\hat{\gamma})$ can be estimated if we fit a model for $E\left[S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|A_i=0,X_i\right]$ using logistic regression in the control arm, with weights $\frac{1}{1-P(A_i=1|X_i;\alpha)}$. It then follows that the second part of 3.17 is in fact the score equation of the intercept in that logistic regression model, which is set to zero through the fitting procedure for γ. As such, the efficient estimator reduces to the following

$$\frac{1}{n}\sum_{i=1}^{n}z(X_i;\hat{\gamma})=\frac{1}{n}\sum_{i=1}^{n}E\left[S\{t,A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|A_i=0,X_i\right] \qquad (3.18)$$

and thus becomes a so-called substitution estimator, which has the advantage of always delivering a result between 0 and 1. While misspecification of the model $E\left[S\{t,A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}|A_i=0,X_i\right]$, the working model for the conditional expectation, does not affect the consistency of $\hat{\theta}$, it does affect its efficiency and may in particular make the estimator less efficient than the RMLE that would be obtained upon setting $d(X_i)=0$. This first term in expression 3.17 for $\hat{\theta}$ is closely related to the mediation formula. It involves averaging the expected outcome values, if the treatment were set to 1, over the mediator distribution if the exposure were set to zero. The second contribution insulates it against bias resulting from possible misspecification of the mediator distribution.

To obtain standard errors for this locally efficient estimator, note that we already have expanded the first contribution

$$\sum_{i=1}^{n} \frac{1-A_i}{1-P(A_i=1|X_i;\alpha)} S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\},$$

which gave us 4 terms: one for the uncertainty in the propensity score U_{i1}, on for the uncertainty in estimating the Cox regression coefficients U_{i2}, one for the cumulative baseline hazard U_{i3}, and finally one for the sample average that we take U_{i4}. In particular, we previously obtained that

$$\sqrt{n}\,\{\hat{S}_{1M_0}(t)-S_{1M_0}(t)\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_{i1}+U_{i2}+U_{i3}+U_{i4}+o_p(1).$$

For the locally efficient estimator, we need to replace U_{i4} by

$$\frac{1-A_i}{1-P(A_i=1|X_i;\alpha)}S\{t|A_i=1,M_i,X_i;\Lambda_0(t,\eta),\eta\}-\theta \\ +\ d(X_i)\,\{A_i-P(A_i=1|X_i;\alpha)\}\,,$$

and since the term $d(X_i)\,\{A_i-P(A_i=1|X_i;\alpha)\}$ also involves the propensity score $P(A_i=1|X_i;\alpha)$, U_{i1} will need to be changed too. We will need to add

$$-d(X_i)\,\{A_i-P(A_i=1|X_i;\alpha)\}$$

to the original term for U_{i1}.

The estimation of $\hat{S}_{0M_0}(t)$ and $\hat{S}_{1M_1}(t)$ follows by simply recoding the exposure. As a result, the restricted maximum likelihood estimators are:

$$\hat{S}_{0M_0}(t)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-A_i}{1-P(A_i=1|X_i;\hat{\alpha})}\,S\{t|A_i=0,M_i,X_i;\hat{\Lambda}_0(t,\hat{\eta}),\hat{\eta}\}\right]$$

and

$$\hat{S}_{1M_1}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i}{P(A_i = 1|X_i;\hat{\alpha})}\,S\{t|A_i = 1, M_i, X_i; \hat{\Lambda}_0(t,\hat{\eta}), \hat{\eta}\}\right].$$

The standard errors are obtained by similarly recoding the exposure. Only for $\hat{S}_{1M_1}(t)$ a $-$ sign needs to be added to U_{1i}. Now that we know the restricted maximum likelihood estimators and its standard errors, we can make them more efficient by solving the following estimation equations:

$$\begin{aligned}0 &= \sum_{i=1}^{n} U_i\{\theta, \alpha, \Lambda_0(t,\eta), \eta\}\\ &= \sum_{i=1}^{n}\frac{1-A_i}{1-P(A_i = 1|X_i;\alpha)}S\{t|A_i = 0, M_i, X_i; \Lambda_0(t,\eta), \eta\} - \theta\\ &\quad + d(X_i)\{A_i - P(A_i = 1|X_i;\alpha)\}\end{aligned}$$

and

$$\begin{aligned}0 &= \sum_{i=1}^{n} U_i\{\theta, \alpha, \Lambda_0(t,\eta), \eta\}\\ &= \sum_{i=1}^{n}\frac{A_i}{P(A_i = 1|X_i;\alpha)}S\{t|A_i = 1, M_i, X_i; \Lambda_0(t,\eta), \eta\} - \theta\\ &\quad + d(X_i)\{A_i - P(A_i = 1|X_i;\alpha)\}\end{aligned}$$

for some index functions $d(X_i)$, where the optimal choices for index functions $d(X_i)$ are obtained by population least squares projection of $(1-A_i) - S\{t|A_i = 0, M_i, X_i; \Lambda_0(t,\eta), \eta\}/\{1-P(A_i = 1|X_i;\alpha)\}$ and $(A_i) - S\{t|A_i = 1, M_i, X_i; \Lambda_0(t,\eta), \eta\}/\{P(A_i = 1|X_i;\alpha)\}$ onto $A_i - P(A_i = 1|X_i;\alpha)$. As a result, $d_{\text{opt}}(X_i)$ equals:

$$\frac{E\left[S\{t|A_i = 0, M_i, X_i; \Lambda_0(t,\eta), \eta\}|A_i = 0, X_i\right]}{1-P(A_i = 1|X_i;\alpha)}$$

for $\hat{S}_{0M_0}(t)$ and

$$-\frac{E\left[S\{t|A_i = 1, M_i, X_i; \Lambda_0(t,\eta), \eta\}|A_i = 1, X_i\right]}{P(A_i = 1|X_i;\alpha)}$$

for $\hat{S}_{1M_1}(t)$.

Suppose now that α, the parameters of the propensity score, and η and $\Lambda_0(t,\eta)$, the nuisance parameters of the survival model, are unknown but substituted by consistent estimators $\hat{\alpha}$, $\hat{\eta}$ and $\hat{\Lambda}_0(t,\eta)$. We can then repeat the previous argument, starting from

$$\begin{aligned}\sqrt{n}\,\{\hat{S}_{1M_0}(t)-S_{1M_0}(t)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(&U_{i1}+U_{i2}+U_{i3}+U_{i4}+\\ &d(X_i)\,\{A_i-P(A_i=1|X_i;\alpha)\}),\end{aligned}$$

while we previously started from

$$\sqrt{n}\,\{\hat{S}_{1M_0}(t)-S_{1M_0}(t)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_{i4}+d(X_i)\,\{A_i-P(A_i=1|X_i;\alpha)\}).$$

The optimal choice of $d(X_i)$ is now obtained by regressing $U_{i1}+U_{i2}+U_{i3}+U_{i4}$ on $A_i-P(A_i=1|X_i;\alpha)$, i.e.

$$d_{\text{opt}}(X_i) = -\frac{E\left[\frac{1-A_i}{1-P(A_i=1|X_i;\alpha)}(U_{1i}+U_{2i}+U_{3i}+U_{4i})\,\{A_i-P(A_i=1|X_i;\alpha)\}\,|X_i\right]}{E\left[\{A_i-P(A_i=1|X_i;\alpha)\}^2\,|X_i\right]}.$$

This will be identical to

$$d_{\text{opt}}(X_i) = -\frac{E\left[\frac{1-A_i}{1-P(A_i=1|X_i;\alpha)}(U_{4i})\,\{A_i-g(X_i;\alpha)\}\,|X_i\right]}{E\left[\{A_i-g(X_i;\alpha)\}^2\,|X_i\right]},$$

which is what we found previously. We can see this because U_{2i} and U_{3i} have mean zero conditional on A_i and X_i and so regressing them on $A_i-P(A_i=1|X_i;\alpha^*)$ gives zero. Since U_{1i} is already of the form $d(X_i)\,\{A_i-P(A_i=1|X_i;\alpha)\}$ for some $d(X_i)$, the residual will eliminate this term.

3.A.2 Eligibility Criteria

A subgroup of the initial population of 1856 breast cancer patients was selected for this study based on several criteria. Those women for whom pCR status was non-identifiable at surgery and the 101 women who progressed beforehand, were excluded from the analysis according to the landmark analysis approach (Giobbie-Hurder et al. 2013). Other selection criteria were: (i) patients who were eligible for the P53 trial (Bonnefoi et al. 2011), (ii) patients who received at least one cycle of neoadjuvant chemotherapy and did not receive radiotherapy before surgery, and (iii) patients without M1, bilateral breast cancer or T4d cancer.

3.A.3 Tables and Figures

Figure 3.3 examines the extent of the efficiency gain of the LE versus the RMLE estimator. The total variance of the estimators equals the variance of the sum of four specific terms. All four terms have to be included, since they each account for a different part of the uncertainty in the estimators. The first term U_{1i} handles the uncertainty due to estimating the propensity score, the second term U_{2i} is included because of the uncertainty in estimating the Cox regression coefficients, the third U_{3i} takes care of the uncertainty due to estimating the cumulative baseline hazard, and the fourth U_{4i} accounts for the fact that we take the sample average. No efficiency gain will be seen for terms U_{2i} and U_{3i}, because we can not improve the estimates of the parameters indexing the Cox regression model. Since the LE estimator relies on covariate information of all patients, and not on that of one treatment group, progress can be made for U_{4i}. Figure 3.3 shows that the total variance $U_{1i}+U_{2i}+U_{3i}+U_{4i}$ of the LE estimator is 30 to 40% smaller over time than that of the RMLE estimator and that this efficiency gain is the result of a smaller variation in $U_{1i}+U_{4i}$.

Figure 3.3: Part of indirect effect variance relative to total indirect effect variance RMLE: efficiency gain in the complete case analysis

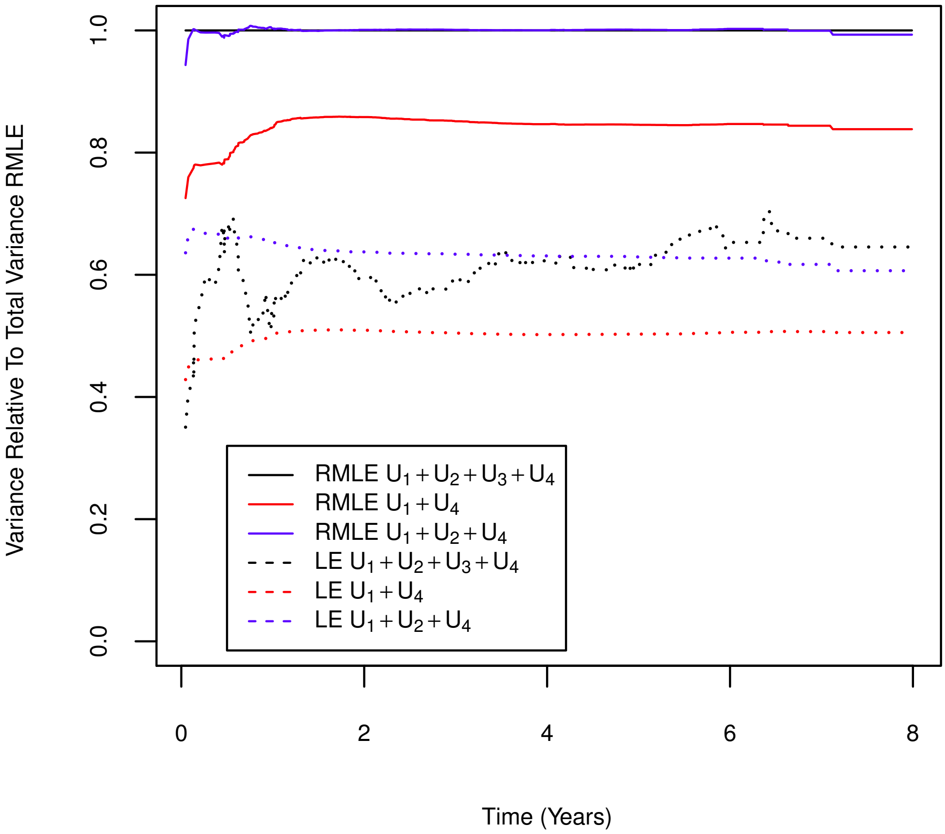


Table 3.3: Model (3.6) at t = 1.010 fitted on the complete cases

		log(OR)	SE	Z-value	P-value
Intercept		4.57	1.27	3.61	0.00
Local PgR status	Negative				
	Positive	0.14	0.68	0.21	0.84
p53	Wild				
	Mutated	-0.38	0.65	-0.58	0.57
Clin. nod. stat.	N0				
	N1	-0.45	0.50	-0.90	0.37
	N2 & N3	0.11	0.86	0.13	0.90
Clin. tum. size	T2				
	T3	-1.13	1.55	-0.73	0.47
	T4	-2.11	2.66	-0.79	0.43
Hist. grade	I				
	II	-0.67	1.21	-0.56	0.58
	III	-1.07	1.42	-0.75	0.45
BC subtype	Lum. A				
	Lum. B (HER2 neg.)	-0.20	1.31	-0.15	0.88
	Lum. B (HER2 pos.)	0.07	1.08	0.07	0.95
	HER2 pos. (non-lum.)	-1.42	1.18	-1.20	0.23
	Triple neg.	0.59	1.30	0.45	0.65
Clin. tum. size x Hist. grade	T2 x I				
	T3 x I				
	T4 x I				
	T2 x II				
	T3 x II	0.96	1.65	0.58	0.56
	T4 x II	1.24	2.80	0.44	0.66
	T2 x III				
	T3 x III	0.94	1.90	0.50	0.62
	T4 x III	2.26	3.07	0.74	0.46
Clin. tum. size x BC subtype	T2 x Lum. A				
	T3 x Lum. A				
	T4 x Lum. A				
	T2 x Lum. B (HER2 neg.)				
	T3 x Lum. B (HER2 neg.)	0.27	1.63	0.17	0.87
	T4 x Lum. B (HER2 neg.)	-0.67	2.49	-0.27	0.79
	T2 x Lum. B (HER2 pos.)				
	T3 x Lum. B (HER2 pos.)	-1.43	1.20	-1.19	0.24
	T4 x Lum. B (HER2 pos.)	-0.95	1.52	-0.62	0.53
	T2 x HER2 pos. (non-lum.)				
	T3 x HER2 pos. (non-lum.)	-0.27	1.42	-0.19	0.85
	T4 x HER2 pos. (non-lum.)	-1.37	1.71	-0.80	0.42
	T2 x Triple neg.				
	T3 x Triple neg.	-1.19	1.33	-0.89	0.37
	T4 x Triple neg.	-2.56	1.69	-1.51	0.13
p53 x BC subtype	Wild x Lum. A				
	Mutated x Lum. A				
	Wild x Lum. B (HER2 neg.)				
	Mutated x Lum. B (HER2 neg.)	0.39	1.17	0.34	0.74
	Wild x Lum. B (HER2 pos.)				
	Mutated x Lum. B (HER2 pos.)	0.41	1.00	0.41	0.68
	Wild x HER2 pos. (non-lum.)				
	Mutated x HER2 pos. (non-lum.)	1.62	1.14	1.42	0.16
	Wild x Triple neg.				
	Mutated x Triple neg.	-0.13	1.09	-0.12	0.90
Clin. nod. stat. x Local PgR status	N0 x Negative				
	N1 x Negative				
	N2 & N3 x Negative				
	N0 x Positive				
	N1 x Positive	0.08	0.76	0.11	0.92
	N2 & N3 x Positive	-1.51	1.29	-1.16	0.24

Table 3.4: Model 3.6 at t = 5.002 fitted on the complete cases

		log(OR)	SE	Z-value	P-value
Intercept		2.52	0.51	4.95	0.00
Local PgR status	Negative				
	Positive	0.17	0.31	0.56	0.58
p53	Wild				
	Mutated	-0.42	0.29	-1.45	0.15
Clin. nod. stat.	N0				
	N1	-0.46	0.26	-1.75	0.08
	N2 & N3	0.25	0.48	0.52	0.60
Clin. tum. size	T2				
	T3	-1.25	0.67	-1.87	0.06
	T4	-2.45	1.49	-1.64	0.10
Hist. grade	I				
	II	-0.73	0.47	-1.56	0.12
	III	-1.08	0.58	-1.87	0.06
BC subtype	Lum. A				
	Lum. B (HER2 neg.)	-0.30	0.59	-0.51	0.61
	Lum. B (HER2 pos.)	-0.01	0.47	-0.02	0.99
	HER2 pos. (non-lum.)	-1.69	0.64	-2.62	0.01
	Triple neg.	0.51	0.57	0.89	0.37
Clin. tum. size x Hist. grade	T2 x I				
	T3 x I				
	T4 x I				
	T2 x II				
	T3 x II	1.06	0.72	1.47	0.14
	T4 x II	1.44	1.56	0.93	0.35
	T2 x III				
	T3 x III	0.93	0.88	1.06	0.29
	T4 x III	2.72	1.80	1.51	0.13
Clin. tum. size x BC subtype	T2 x Lum. A				
	T3 x Lum. A				
	T4 x Lum. A				
	T2 x Lum. B (HER2 neg.)				
	T3 x Lum. B (HER2 neg.)	0.40	0.78	0.51	0.61
	T4 x Lum. B (HER2 neg.)	-0.90	1.40	-0.64	0.52
	T2 x Lum. B (HER2 pos.)				
	T3 x Lum. B (HER2 pos.)	-1.67	0.56	-2.99	0.00
	T4 x Lum. B (HER2 pos.)	-1.11	0.86	-1.29	0.20
	T2 x HER2 pos. (non-lum.)				
	T3 x HER2 pos. (non-lum.)	-0.20	0.76	-0.27	0.79
	T4 x HER2 pos. (non-lum.)	-1.91	1.36	-1.40	0.16
	T2 x Triple neg.				
	T3 x Triple neg.	-1.34	0.62	-2.17	0.03
	T4 x Triple neg.	-3.31	1.34	-2.47	0.01
p53 x BC subtype	Wild x Lum. A				
	Mutated x Lum. A				
	Wild x Lum. B (HER2 neg.)				
	Mutated x Lum. B (HER2 neg.)	0.45	0.53	0.85	0.39
	Wild x Lum. B (HER2 pos.)				
	Mutated x Lum. B (HER2 pos.)	0.53	0.50	1.06	0.29
	Wild x HER2 pos. (non-lum.)				
	Mutated x HER2 pos. (non-lum.)	1.83	0.65	2.82	0.00
	Wild x Triple neg.				
	Mutated x Triple neg.	-0.04	0.54	-0.08	0.94
Clin. nod. stat. x Local PgR status	N0 x Negative				
	N1 x Negative				
	N2 & N3 x Negative				
	N0 x Positive				
	N1 x Positive	0.05	0.36	0.14	0.89
	N2 & N3 x Positive	-1.91	0.77	-2.50	0.01

3.A.4 Sensitivity analysis one: Analysis on the complete subset of eligible patients using the missing-indicator method

The RMLE and LE estimator are used to estimate the indirect effect of chemotherapy arm on DFS mediated via pCR and the remaining direct effect. The multivariate Cox regression model presented in Table 3.5 served as outcome model for the first step of the RMLE and LE estimator in the analysis on the complete subset of eligible patients. To obtain the LE estimator, we fit an additional logistic regression model for each time t separately. We used the same predictors as in the outcome model (Table 3.5), except those involving pCR and treatment. Figure 3.4 presents the direct effect ratio of chemotherapy on DFS on the right-hand side and the indirect effect ratio via pCR on the left-hand side. The RMLE estimator yields a direct effect of 1.115 (95% CI 1.045 to 1.184) and an indirect effect of 1.002 (95% CI 0.975 to 1.030) after 5 years. The LE estimator gave a similar direct effect of 1.115 (95% CI 1.042 to 1.188) and a more precise indirect effect of 1.000 (95% CI 0.979 to 1.021). On the basis of this, we may conclude that the probability that the duration of DFS lasts longer than 5 years after administering the experimental taxane-based regimen is about 11.7% and 11.5% larger for the RMLE and LE estimator respectively than when the anthracycline based regimen would be administered. A very small part of the total intention-to-treat effect is due to the effect via pCR. In particular, the mediation proportion shows that only 0.3% of the treatment effect on the DFS risk difference is mediated by the treatment effect on pCR for the LE estimator after 5 years (Figure 3.5).

Figure 3.6 examines the extent of the efficiency gain of the LE versus the RMLE estimator in the first sensitivity analysis. It shows that the total variance $U_{1i} + U_{2i} + U_{3i} + U_{4i}$ of the LE estimator is 30 to 40% smaller over time than that of the RMLE estimator and that this efficiency gain is the result of a smaller variation in $U_{1i} + U_{4i}$.

Figure 3.4: Results of the first sensitivity analysis: direct and indirect effect risk ratios of surviving the given time indicated on the X-axis with accompanying 95% point-wise confidence intervals

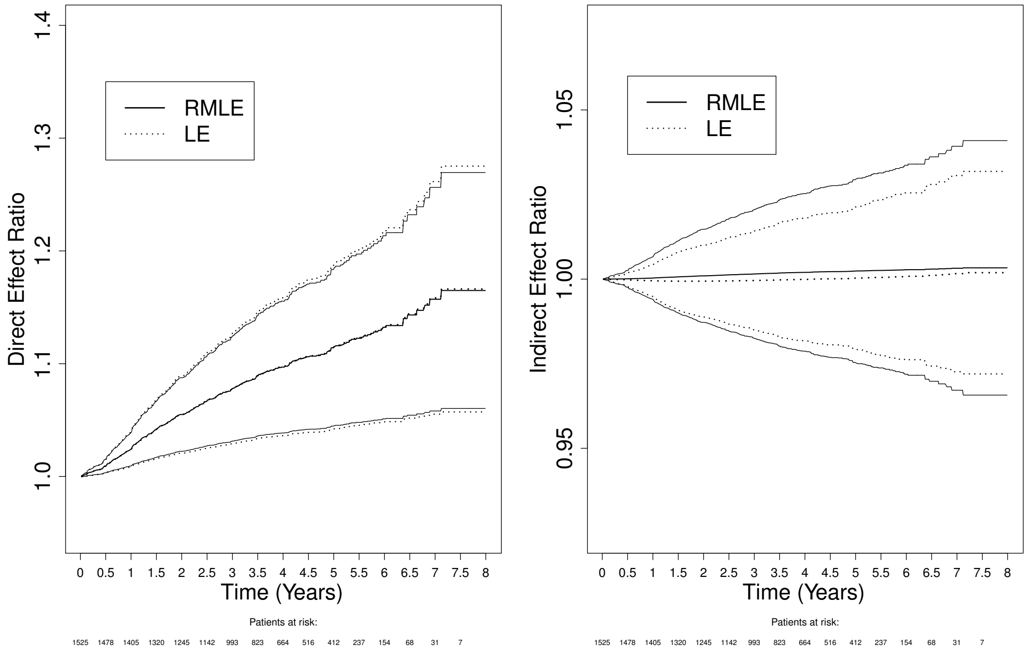


Figure 3.5: Results of the first sensitivity analysis: proportion mediated for the given time indicated on the X-axis

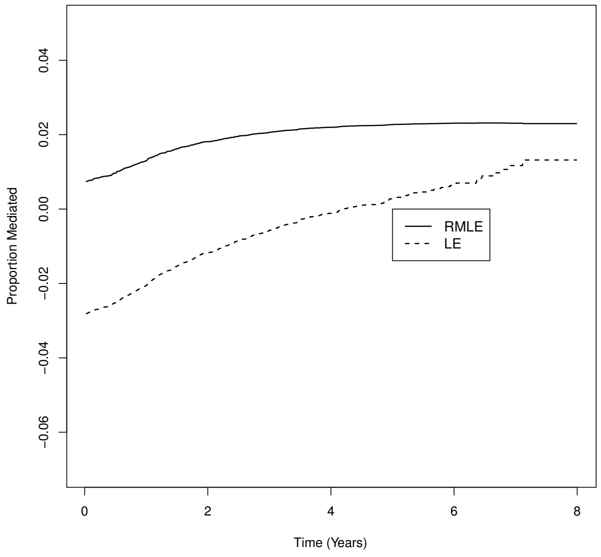


Figure 3.6: Part of indirect effect variance relative to total indirect effect variance RMLE: efficiency gain in the first sensitivity analysis

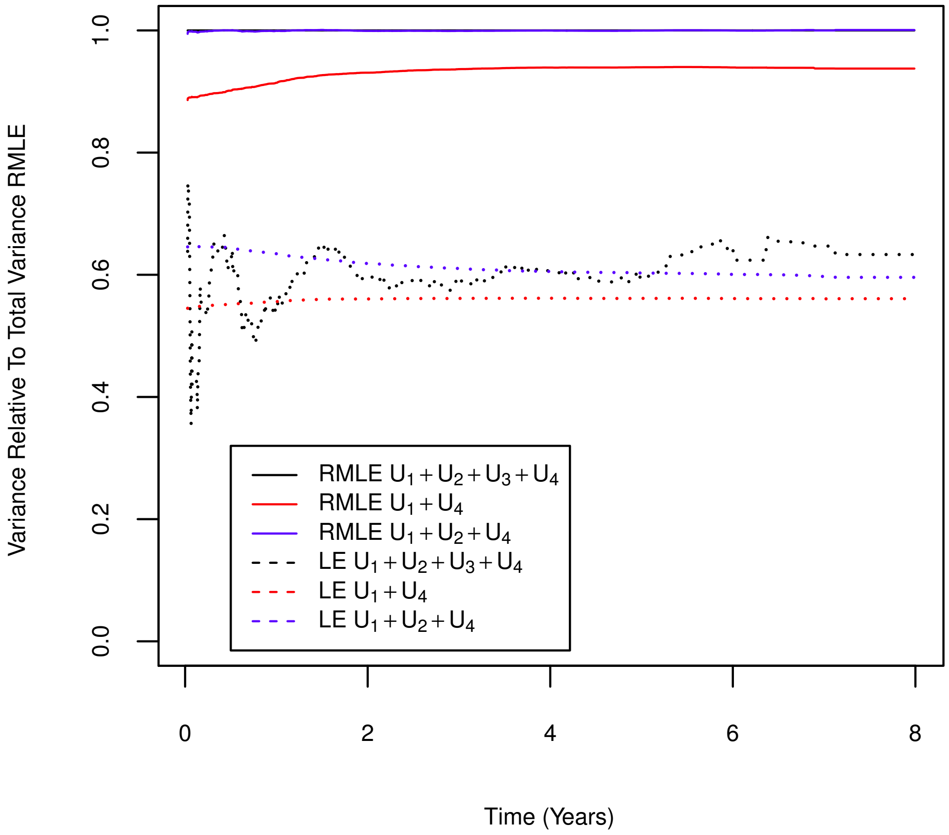


Table 3.5: Cox regression outcome model of the first sensitivity analysis

		HR	95% CI	P-value
Treatment	Standard			
	Experimental	0.72	0.60 - 0.87	0.00
pCR	No			
	Yes	0.80	0.52 - 1.21	0.28
p53	Wild			
	Mutated	1.10	0.88 - 1.38	0.38
	Missing	0.78	0.59 - 1.03	0.08
Local ER status	Negative			
	Positive	0.53	0.36 - 0.77	0.00
	Unknown	0.43	0.22 - 0.83	0.01
Clin. nod. stat.	N0			
	N1	1.89	1.50 - 2.37	0.00
	N2 & N3	2.31	1.54 - 3.48	0.00
Clin. tum. size	T2			
	T3	3.21	1.21 - 8.47	0.02
	T4	4.15	0.85 - 20.22	0.08
Hist. grade	I			
	II	2.14	1.03 - 4.45	0.04
	III	2.36	1.03 - 5.37	0.04
	Missing	1.30	0.50 - 3.36	0.59
BC subtype	Lum. A			
	Lum. B (HER2 neg.)	1.63	0.85 - 3.12	0.14
	Lum. B (HER2 pos.)	1.28	0.74 - 2.22	0.37
	HER2 pos. (non-lum.)	1.69	0.88 - 3.27	0.12
	Triple neg.	0.56	0.28 - 1.10	0.09
	Missing	1.22	0.73 - 2.03	0.45
Hist. type	Inv. ductal			
	Inv. lobular	1.01	0.74 - 1.39	0.94
	Other	1.05	0.68 - 1.63	0.81
Clin. nod. stat. x pCR	N0 x No			
	N1 x No			
	N2 & N3 x No			
	N0 x Yes			
	N1 x Yes	0.30	0.16 - 0.57	0.00
	N2 & N3 x Yes	0.20	0.05 - 0.91	0.04
Clin. tum. size x Hist. grade	T2 x I			
	T3 x I			
	T4 x I			
	T2 x II			
	T3 x II	0.33	0.12 - 0.92	0.03
	T4 x II	0.58	0.11 - 2.90	0.50
	T2 x III			
	T3 x III	0.49	0.16 - 1.52	0.22
	T4 x III	0.29	0.05 - 1.67	0.17
	T2 x Missing			
	T3 x Missing	0.55	0.16 - 1.93	0.35
	T4 x Missing	0.72	0.12 - 4.35	0.72

Table 3.5: Cox regression outcome model of the first sensitivity analysis

		HR	95% CI	P-value
Clin. tum. size x BC subtype	T2 x Lum. A			
	T3 x Lum. A			
	T4 x Lum. A			
	T2 x Lum. B (HER2 neg.)			
	T3 x Lum. B (HER2 neg.)	0.56	0.21 - 1.50	0.25
	T4 x Lum. B (HER2 neg.)	1.41	0.36 - 5.49	0.62
	T2 x Lum. B (HER2 pos.)			
	T3 x Lum. B (HER2 pos.)	2.31	1.11 - 4.78	0.02
	T4 x Lum. B (HER2 pos.)	1.49	0.60 - 3.68	0.39
	T2 x HER2 pos. (non-lum.)			
	T3 x HER2 pos. (non-lum.)	0.74	0.31 - 1.75	0.49
	T4 x HER2 pos. (non-lum.)	1.02	0.35 - 3.01	0.97
	T2 x Triple neg.			
	T3 x Triple neg.	3.06	1.39 - 6.75	0.01
	T4 x Triple neg.	5.90	2.05 - 17.00	0.00
	T2 x Missing			
	T3 x Missing	1.56	0.76 - 3.21	0.23
	T4 x Missing	0.89	0.36 - 2.16	0.79

3.A.5 Sensitivity analysis two: Smaller number of baseline covariates

Finally, we report the results of an additional mediation analysis that makes use of an outcome model that included a smaller number of baseline covariates. This outcome model excluded the important, but incomplete baseline covariates local PgR status, p53 status, histological grade and intrinsic breast cancer subtype. The multivariate Cox regression model is presented in Table 3.6. An additional logistic regression model for each time t separately is used to obtain the LE estimator. This model consists of the same predictors as the outcome model (Table 3.6), except those involving pCR and treatment. Figure 3.7 presents the direct effect ratio of chemotherapy on DFS on the right-hand side and the indirect effect ratio via pCR on the left-hand side. The RMLE estimator yields a direct effect of 1.114 (95% CI 1.014 to 1.214) and an indirect effect of 1.006 (95% CI 0.979 to 1.033) after 5 years. The LE estimator gave a similar direct effect of 1.114 (95% CI 1.012 to 1.216) and a more precise indirect effect of 1.005 (95% CI 0.983 to 1.027). Thus, we may conclude that the probability that the duration of DFS lasts longer than 5 years after administering the experimental taxane-based regimen is about 12.1% and 12% larger for the RMLE and LE estimator respectively than when the anthracycline based regimen would be administered. A very small part of the total intention-to-treat effect is due to the effect via pCR. In particular, the mediation proportion shows that only 4.5% of the treatment effect on the DFS risk difference is mediated by the treatment effect on pCR for the LE estimator after 5 years (Figure 3.8).

Figure 3.9 again examines the extent of the efficiency gain of the LE versus the RMLE estimator. It shows that the total variance $U_{1i}+U_{2i}+U_{3i}+U_{4i}$ of the LE estimator is 30 to 50% smaller over time than that of the RMLE estimator and that this efficiency gain is the result of a smaller variation in $U_{1i}+U_{4i}$.

Figure 3.7: Results of the second sensitivity analysis: direct and indirect effect risk ratios of surviving the given time indicated on the X-axis with accompanying 95% point-wise confidence intervals

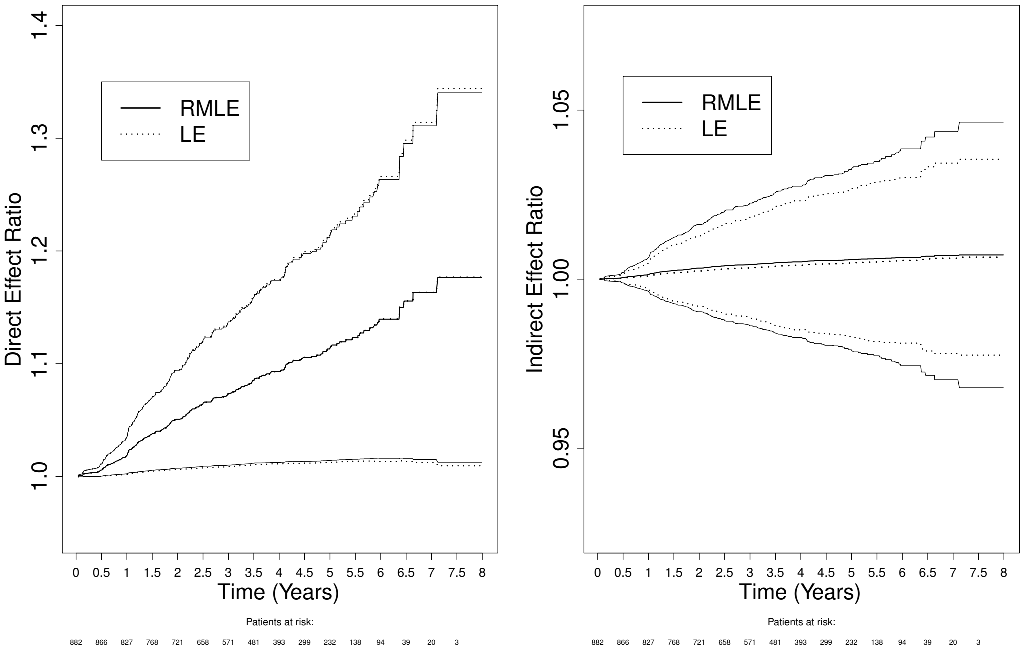


Figure 3.8: Results of the second sensitivity analysis: proportion mediated for the given time indicated on the X-axis

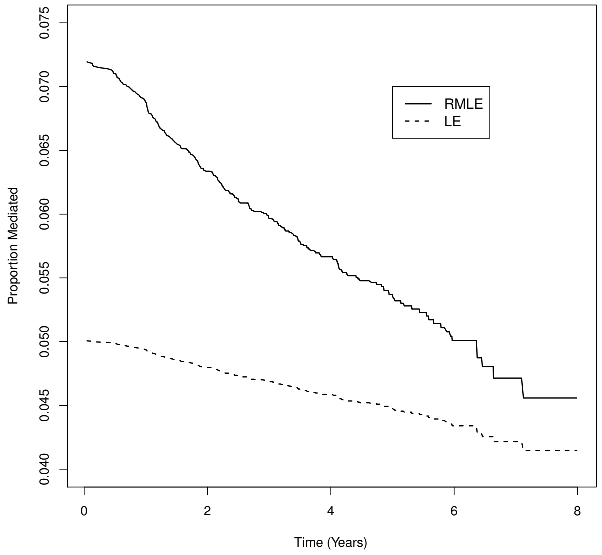


Figure 3.9: Part of indirect effect variance relative to total indirect effect variance RMLE: efficiency gain in the second sensitivity analysis

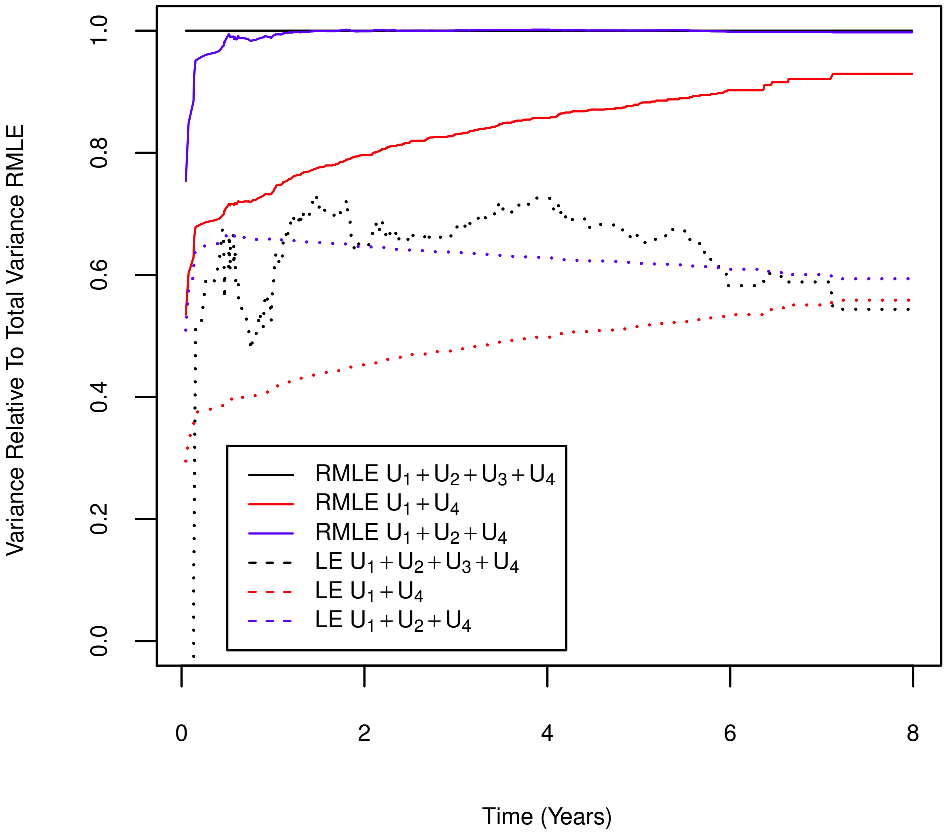


Table 3.6: Cox regression outcome model of the second sensitivity analysis

		HR	95% CI	P-value
Treatment	Standard			
	Experimental	0.74	0.58 - 0.95	0.02
pCR	No			
	Yes	0.99	0.59 - 1.67	0.98
Clin. nod. stat.	N0			
	N1	1.70	1.28 - 2.25	0.00
	N2 & N3	2.70	1.49 - 4.91	0.00
Clin. tum. size	T2			
	T3	1.93	1.48 - 2.52	0.00
	T4	2.91	2.04 - 4.15	0.00
Clin. nod. stat. x pCR	N0 x No			
	N1 x No			
	N2 & N3 x No			
	N0 x Yes			
	N1 x Yes	0.40	0.18 - 0.89	0.02
	N2 & N3 x Yes	0.16	0.02 - 1.29	0.09

3.A.6 Simulations

We reported large sample theory in this Appendix in section 3.A.1. To investigate the finite sample properties of the estimators and their inference some simulations were done. We evaluate the performance of our two estimators through a simulation analysis with 1000 runs for data sets of 100, 200, 500 and 1000 observations. First, a dichotomous exposure A is drawn with $P(A=0)=P(A=1)=0.5$. For each simulation, we report the results at 3 different time points (3, 5 and 8 years). Sandwich standard errors (see the Appendix) are presented and used to construct 95% confidence intervals. Covariates $X=(X_1,X_2)^t$ are generated as follows: X_1 and $X_2 \sim \mathcal{N}(0,0.49)$. The dichotomous mediator M is generated as a Bernoulli variate obeying a $\text{logit}\{P(M=1|A,X)\}=\alpha_0+\alpha_1 A+\alpha_2^T X^*$ with X^* including the 2 covariates of X and their interaction. Parameter values were set to $\alpha_0 = 0$, $\alpha_1=0.5$ and α_2^T equals $(-0.25,0.15,-0.2)^t$. Finally, the event time T is drawn from a Weibull distribution with shape parameter $a=1$ and scale parameter $b=1/\{\lambda_T \exp(\beta_1 A+\beta_2 M+\beta_3^T X^*)\}$ with $\lambda_T=0.2$, $\beta_1=-0.4$, $\beta_2=0.8$ and $\beta_3^T=(-0.4,-0.4,0.2)^t$. The censoring time C is also drawn from a Weibull distribution with shape parameter $a=1$ and scale parameter $b=1/\lambda_C=0.3$. An event occurs if $T^*=\min(T,C)$ equals T. Results of the simulation analyses with data sets of 100, 200, 500 and 1000 observations are presented in Tables 3.7, 3.8, 3.9 and 3.10 respectively.

Table 3.7: Results simulations analysis $n=100$

	RMLE RR$_\text{D}$			LE RR$_\text{D}$			RMLE RR$_\text{M}$			LE RR$_\text{M}$		
	3	5	8	3	5	8	3	5	8	3	5	8
Bias	0.04	0.09	0.41	0.03	0.09	0.40	0.01	0.01	0.03	0.00	0.00	0.01
Emp. Sd	0.32	0.63	4.16	0.32	0.63	4.28	0.08	0.12	0.18	0.06	0.08	0.12
Se	0.48	0.91	2.18	0.51	1.00	2.64	0.08	0.11	0.157	0.07	NaN	NaN
Coverage	0.98	0.97	0.95	0.98	0.98	0.96	0.95	0.92	0.91	0.98	NaN	NaN

Table 3.7, 3.8, 3.9 and 3.10 show that the proposed efficient estimators LE delivers drastic efficiency gains for the natural indirect effect, although not for the natural direct effect in comparison to the RMLE estimator. This is not surprising because the choice whether or not to include baseline covariates only affects estimation of the indirect effect (Vandenberghe et al. 2017a). Both estimators are generally unbiased, except in the simulation study with only 100 observations, where some

Table 3.8: Results simulations analysis $n = 200$

	RMLE RR_D			LE RR_D			RMLE RR_M			LE RR_M		
	3	5	8	3	5	8	3	5	8	3	5	8
Bias	0.01	0.03	0.09	0.01	0.03	0.09	0.00	0.00	0.01	0.00	0.00	0.01
Emp. Sd	0.20	0.33	0.65	0.20	0.33	0.65	0.05	0.08	0.11	0.04	0.06	0.08
Se	0.22	0.41	0.63	0.24	0.44	0.70	0.06	0.08	0.11	0.05	0.07	0.08
Coverage	0.97	0.97	0.92	0.97	0.98	0.93	0.97	0.97	0.95	0.99	0.98	0.94

Table 3.9: Results simulations analysis $n = 500$

	RMLE RR_D			LE RR_D			RMLE RR_M			LE RR_M		
	3	5	8	3	5	8	3	5	8	3	5	8
Bias	0.00	0.01	0.03	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00
Emp. Sd	0.12	0.22	0.37	0.12	0.22	0.37	0.03	0.05	0.06	0.02	0.03	0.05
Se	0.12	0.17	0.34	0.12	0.18	0.34	0.04	0.05	0.09	0.03	0.04	0.07
Coverage	0.94	0.90	0.93	0.94	0.90	0.93	0.98	0.98	0.99	0.99	0.98	0.99

Table 3.10: Results simulations analysis $n = 1000$

	RMLE RR_D			LE RR_D			RMLE RR_M			LE RR_M		
	3	5	8	3	5	8	3	5	8	3	5	8
Bias	0.00	0.01	0.02	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Emp. Sd	0.08	0.14	0.24	0.08	0.14	0.24	0.02	0.03	0.05	0.02	0.02	0.04
Se	0.08	0.14	0.20	0.08	0.14	0.21	0.03	0.04	0.06	0.02	0.03	0.05
Coverage	0.95	0.94	0.91	0.96	0.95	0.91	0.98	0.99	0.99	0.99	0.99	0.99

bias appears due to the small sample size. This is especially visible at the latest time point, where there's not much information left to fit the model. The standard errors also become larger due to lack of information.

CHAPTER 4

4

Longitudinal mediation analysis in a survival setting

4.1 Introduction

4.2 Setting

4.3 Proposal

4.4 Extensions

CHAPTER 5

Conclusion and Future Research

5.1 Conclusion

Over recent years there has been a growing awareness of the usefulness of mediation analyses to supplement the intention-to-treat analysis in randomised experiments. Applied researchers are increasingly using (natural) direct and indirect effects as they try to grasp the working mechanism of a certain treatment. In this thesis, we briefly discussed several research areas which could benefit from supplementing the usual intention-to-treat analysis with a mediation analysis. We saw two areas of research wherein especially the (natural) indirect effect could be of interest: (1) the development of better, more efficient treatments with fewer side effects (Marso et al. 2016) and (2) the evaluation of putative surrogate markers that could potentially replace long-term more clinically relevant outcomes (Vandenberghe et al. In press). Additionally, two examples were discussed which showed that also the (natural) direct effect could add helpful information to the analyses of randomised trials: (1) in assessing the usefulness of existing treatments for problems or diseases for which they were not originally developed (McIntyre et al. 2014) and (2) in clarifying seemingly ambiguous results on total intention-to-treat effects (Padian et al. 2007).

The traditional approach to direct and indirect effects dates back to the social science literature of the 1970s and early 1980s (Baron and Kenny 1986; Judd and Kenny 1981) and can be viewed as a refining of earlier ideas from Wrights' path analysis (Wright 1920). A number of papers have been published that show how using these traditional approaches, as the difference-of-coefficients and the product-of-coefficients method, may produce misleading results for non-linear models (Robins and Greenland 1992; Kaufman et al. 2004; Ten Have and Joffe 2012; VanderWeele 2011). The proposals of Robins and Greenland (1992) and Pearl (2001) delivered the possibility of a formal mediation analysis framework based on counterfactual-based distribution-free definitions of natural direct and indirect effects that can be identified under certain well-defined conditions. Pearl's *mediation formula* (Pearl 2001, 2012) enables the combination of arbitrary statistical models for the outcome and mediator to identify natural direct and indirect effects and is thereby far superior to the traditional approach to direct and indirect effects.

Although this identification result was groundbreaking for modern causal mediation analysis, it still has a number of important limitations (Vansteelandt et al. 2012a). A first important limitation, that recently fueled an ongoing debate on the usefulness of natural direct and indirect effects in the epidemiologic literature, is the need for cross-world assumptions to be able to identify natural direct and indirect effects. This is indeed a very strong assumption as it can not be empirically verified or guaranteed by study design (Robins and Greenland 1992). Second, applying the mediation formula to compute natural direct and indirect effects involves integration which can become very complex for certain combinations of mediator and outcome models. Third, combining arbitrary models for the mediator and outcome may result in complex closed-form expressions of natural direct and indirect effects, if they can be defined altogether (Lange et al. 2012). VanderWeele and Vansteelandt (2010) show that even a simple linear model for the mediator and a logistic regression model for the outcome already result in elaborate expressions for the natural direct and indirect effect. A final limitation, that may be considered a major disadvantage in particular in the analysis of randomised trials as the concern for model misspecification is more pertinent in that setting, is that this approach may deliver biased effect estimates if the model for the mediator or the model for the outcome (or both) are misspecified. Although van der Laan and Petersen (2008) too need to rely on cross-world assumptions, they accommodate the other concerns by directly modeling the natural direct effect of interest via so-called double robust estimators that require correct specification of a model for the distribution of the mediator, and either for the distribution of the exposure or the mean outcome. Tchetgen Tchetgen and Shpitser (2012) and Tchetgen Tchetgen (2011) additionally discuss estimation of natural indirect effects via their triple-robust estimators that require 2 out of the 3 models (one for exposure, mediator and outcome) to be correctly specified. These alternatives still involve inverse weighting with the mediator distribution however, which can sometimes yield erratic behaving estimators that tend to be sensitive to minor misspecifications in the tails of the mediator distribution (Vansteelandt 2012b). In **Chapter 2** and **Chapter 3**, we therefore propose semi-parametric efficient strategies that exploit our a priori knowledge of the randomisation probabilities and as a result do not rely on a model for the mediator to obtain unbiased estimates for natural direct and indirect effects.

In **Chapter 2**, we first discuss a simple estimator for natural direct and indirect effects that delivers unbiased estimates for binary and continuous outcomes by only relying on the correct specification of the model for the outcome. However, when baseline covariates that are predictive of the mediator are available, not relying on a model for the mediator distribution is far less efficient than maximum likelihood estimation (MLE) based on the mediation formula. Therefore, we proposed two estimators that, like MLEs, make use of a model for the mediator to improve efficiency by optimally extracting information from baseline covariate data, but are robust against misspecification, unlike MLEs. In particular, we propose the locally efficient (LE) estimator that is efficient when a working model for the mediator is correctly specified and the restricted efficient (RE) estimator that achieves efficiency within a more restrictive class of estimators regardless of correct specification of a working model for the mediator. As simulation studies showed similar behavior for both estimators, we recommend to use the LE estimator in practice as it is relatively simple to use in routine applications. These two proposed estimators have several advantages as compared to the popular MLEs and robust estimators of

Tchetgen Tchetgen and Shpitser (2012) and Lendle et al. (2013). First, unlike those proposals it does not require inverse weighting with the mediator distribution. In fact, when the randomisation probabilities are constant, it does not require inverse probability weighting at all. Second, in our experience models for the distribution of the mediator (or the probability of treatment, given mediator and covariates) are generally more difficult to specify than models for the mean of the outcome. And finally our approach is much more easy to apply because of its greater simplicity.

In **Chapter 3**, the strategy of the locally efficient estimator is further extended to handle time-to-event outcomes. This proposal has the advantage of quantifying the natural direct and indirect effect on the survival scale, instead of the hazard scale, as in earlier work on mediation analysis with time-to-event outcomes from Lange and Hansen (2011) and VanderWeele (2011). This improves interpretability since hazard ratios lack causal interpretation, even in randomised trials (Hernan 2010; Aalen et al. 2015), but does require a more high-dimensional graphical representation of the results. Although this approach can be generally applied to mediation analyses of time-to-event outcomes, this chapter was written as an example of how mediation analyses may be informative to evaluate the appropriateness of potential intermediate outcomes or surrogate markers when data from just a single trial are available. Even though mediation is not a necessary condition for a mediator to be a good surrogate marker (Joffe and Greene 2009; VanderWeele and Vansteelandt 2013), we show that mediation analysis techniques may be informative in that respect.

Finally, because focussing on a single mediator measured at a single point in time is rather limiting, recent advances in mediation analysis with multiple or repeatedly measured mediators are discussed in **Chapter 4**. Traditional approaches to mediation analysis with multiple mediators are criticized because the literature is rather unclear about the interpretations of the effects and the assumptions needed for them to be identified (VanderWeele and Vansteelandt 2009, 2010; Imai et al. 2010). The mediation analysis strategy that we propose for a randomly assigned exposure, repeatedly measured mediator and time-to-event endpoint has several advantages compared to the traditional approach, the so-called dynamic path analysis (Strohmaier et al. 2015; Fosen et al. 2006). When an additive hazards model is used for the time-to-event outcome in combination with a linear regression model for the normal mediator with constant variance, the resulting direct and indirect effects are comparable. Our approach however is not restricted to this combination of models and can handle different types of survival models (i.e. from additive hazards to proportional hazards models) and different kinds of mediators. As it also allows long term effects of covariates and mediators on covariates, mediators and the outcome measured later in time, the presence of time-varying confounders and unmeasured confounders of different realizations of the mediator over time, it can be applied to much more realistic settings than the traditional approach. Our proposal is closely related to recent contributions of Zheng and van der Laan (2012b) and Zheng and van der Laan (2017), but there are subtle differences that will be discussed in detail in Section 5.3.2.

5.2 On imperfections of the current literature

As recognition of the usefulness of mediation analyses grows, an increasing number of applied papers are published wherein direct and indirect effects are being used to better understand the treatment mechanism. All too often however the seemingly intuitive and appealingly simple regression procedure, known as the difference-of-coefficients method, is used (Naimi 2015). First, a simple regression model of the outcome regressed against the exposure (and possibly baseline covariates) is fitted. In a second step, it seems intuitive to further adjust for the mediator to block that part of the exposure effect going through said mediator. The remaining effect of treatment is then interpreted as the direct effect, and the change in exposure effect is interpreted as indirect effect. Alternatively, the indirect effect is calculated as the product of the exposure effect on the mediator and the mediator effect on the outcome. This approach, better known as the product-of-coefficients method, equally lacks a theoretical basis beyond linear models, but at least has the advantage of being useful as a test of the null hypothesis of no indirect effect (VanderWeele 2011; Vansteelandt et al. 2012a). Although numerous papers have been published that discuss how these traditional approaches to direct and indirect effects are only valid in the absence of unit-level interactions and with linear regression models (Robins and Greenland 1992; Kaufman et al. 2004; Ten Have and Joffe 2012; VanderWeele 2011), a number of recent applied papers that discuss results about direct and indirect effects still make use of these traditional methods (Contiero et al. 2013; Wang et al. 2010; Bränström et al. 2010). This phenomenon of continually using methods that are proven to be limited to linear models, leads to the conclusion that the literature on the topic of causal mediation analysis is seriously lacking. In particular, the literature is mainly directed to highly technical audiences in statistics. The focus lies too much on counterfactual-based definitions, assumptions, estimators and influence functions and offers too little practical guidance for applied researchers through concrete examples of how to implement these estimators. With our paper about mediation analyses of time-to-event outcomes (Vandenberghe et al. In press), we tried to contribute by providing a step-by-step explanation of how to implement the simple estimator proposed by Tchetgen Tchetgen (2011) and our own proposal, which in our opinion really clarified the difference between the two and the advantage of the locally efficient estimator. Additionally, an easy-to-use R function was provided, but applied researchers would really benefit from a comprehensive software package including different estimators together with a detailed and easy to understand discussion on their relative advantages and disadvantages. As such, the still existing barrier to routine application of these methods might be gradually broken down.

Although more and more applied papers are being published on the topic of mediation analysis, the usefulness of the natural direct and indirect effect has been the subject of a recent debate in the epidemiologic literature (Naimi et al. 2014). Their relevance has been questioned as they do not directly correspond to real-life interventions that might help improve population health. The causal inference literature has generally argued that causal effects cannot be estimated unless a clear (possibly hypothetical) intervention can be defined (Glass et al. 2013; Hernan 2005). As a result, some authors (Robins and Greenland 1992; Kaufman 2009) have expressed their concerns about the impossibility to conduct experiments in which the identification assumptions for natural direct and indirect effects are satisfied and these effects could thus be estimated. This apprehension originates from the fact that natural direct and indirect effects are defined in terms of so-called cross-world counterfactuals that are unobservable, even from randomised controlled trials. Others (Pearl 2001; Schwartz et al. 2010) defend the use of natural direct and indirect effects as they can be identified under specific assumptions and provide useful information about existing mechanisms. Moreover, randomised experiments very often cannot be conducted due to practical or ethical considerations and results of causal inference should thus not be discarded as they provide a means of getting closer to the 'truth' where other options fail. Naimi et al. (2014) cite Joffe et al. (2001) to argue that natural effects are irrelevant in public health research, in contrast to controlled direct effects, as they do not connect to the effect of particular policies. Naimi et al. (2014) revisit the example of breast cancer risk for women taking hormone replacement therapy. More breast cancers are observed among women taking hormone replacement therapy, but as these women are also subject to more mammographic screening, the question remains how much of the excess cases are due to increased detection. In this example, a natural direct effect would provide us with information about the breast cancer risk under no postmenopausal hormone therapy and how this risk would change for women under postmenopausal hormone therapy, had mammographic screening not been affected by the therapy. We can agree that it is hard to think about an intervention that results in exposed women undergoing the mammography screening they would have undergone had they not been exposed (and vice versa). As such, a controlled direct effect that tells us the difference in breast cancer risk under postmenopausal hormone therapy and no therapy if all women would have been screened indeed seems more relevant. However, we can think about examples that show the relevance of natural effects over controlled direct effects as well. Remember the MIRA trial (Padian et al. 2007; Rosenblum et al. 2009), which we used to describe the setting where mediation analysis could help clarify seemingly ambiguous results on total intention-to-treat effects. In this setting, the natural direct effect would tell us about the HIV risk under standard treatment and how this risk would change for patients that additionally received diaphragm and lubricant gel had this not had an effect on condom use. This is clearly a far more relevant question than what the controlled direct effect would provide an answer to: the change in HIV risk under standard treatment versus additional diaphragm and lubricant gel availability if no one (or everyone) used condoms. We definitely recognize the usefulness of relatively non-technical papers about natural effects and their identification assumptions like the ones of Naimi et al. (2014) and Hafeman and Schwartz (2009), anti and pro natural effects respectively. But as Kaufman (2009) criticizes Hafeman and Schwartz (2009) for focussing exclusively on natural direct and indirect effects and forgetting about the usefulness of controlled direct effects, their discussion too could certainly be more nuanced and provide concrete examples when natural (in)direct effects versus controlled direct effects are more relevant.

5.3 Future research

5.3.1 Meta-analysis

The meta-analytic approach is still considered to be the gold standard approach for the validation of surrogate markers (Daniels and Hughes 1997; Buyse et al. 2000a; Alonso et al. 2015). Although mediation analysis techniques can be informative to study potential surrogate markers and have the advantage over the meta-analysis approach of needing only a single trial, as they depart from limited information, they do not have the capacity to examine between-study heterogeneity and the results are thus less well generalisable than results from a meta-analysis. In order to generalise mediation analysis results, the direct treatment effect on the outcome not mediated by the surrogate would have to be fairly similar across studies, otherwise knowledge about the treatment effect on the surrogate would not allow prediction of the treatment effect on the outcome (if there is also a relatively stable causal relationship between the surrogate and the outcome). This quantity of predictive ability is measured via the trial-level R^2 under the meta-analytic framework. The meta-analysis approach on the other hand has the disadvantage, that its power to detect a high trial-level R^2 may be very weak when there is little variability in the effect of treatment on the surrogate across trials (Joffe and Greene 2009). This could potentially happen when data from the different trials are homogenous, although this does not mean that the potential surrogate is a poor one, however. Recent contributions to the causal inference literature discuss issues of so-called ‘transportability’ and the potential to inform decisions in similar and in other populations as the effects were estimated (Hernan and VanderWeele 2011). Thus, although it is quite restrictive that the direct treatment effect on the outcome would have to be fairly similar across studies for the results of the mediation analysis to be generalisable, future research could for instance examine if the direct effect is similar for different patient groups. One could for instance fit the Cox regression model and the logistic regression model on the predictions from that Cox model as in **Section 3.4** of **Chapter 3** on the EORTC 10994/BIG 1-00 randomised phase 3 trial data and use this as a *training* data set, but take the average from the final step across patients from a new data set where our baseline covariates were also observed (i.e. *test* data set). As such, one could examine if the direct effect is indeed similar across different populations.

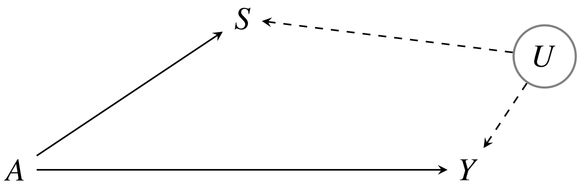


Figure 5.1: Causal diagram representation of surrogacy.

A disadvantage of both the mediation analysis and the meta-analysis in the evaluation of surrogate markers is that neither the proportion mediated, nor the trial-level R^2 is able to provide information about the expected treatment effect of a new treatment. If the mediation analysis results indicate however that their is an indirect effect and thus a causal effect of the surrogate on the outcome then, if we adequately controlled for confounding of their relationship, it is reasonable to assume that this new treatment will have an effect on the outcome of interest if the data show that it has an effect on the surrogate. VanderWeele (2013) criticizes the mediation analysis approach to evaluate surrogacy because it will not detect a surrogate unless the surrogate has a direct causal effect on the outcome of interest. It is argued that the surrogate might still be valuable in that case because of common causes U of the surrogate and the outcome (as in Figure 5.1). Now imagine an example as in Figure 5.1 where surrogate S has no direct effect on outcome Y, but due to unmeasured common cause U, say age, S serves as a very good proxy for Y: S and Y occur much more often in older patients. Then if a future study (unknowingly) selects patients with little variation in terms of age, S will (surprisingly) be a very poor surrogate for Y although previous studies showed different results. In our opinion, evidence of mediation will thus generally add support to the validity of a candidate surrogate. To conclude, it is rather difficult to choose one approach over the other at this moment because they both have appealing properties, but also disadvantages. Future research should take away this uncertainty by for instance examining if meta-analytic proven surrogates are also supported by mediation analysis results (i.e. CD4 count). It would be informative to see specific settings where and the reason why the two approaches may give contradictory results. It would even be more valuable to not have to choose at all if future research would find ways of integrating these two approaches to evaluate surrogacy.

5.3.2 Immortal time bias

In our paper about mediation analyses of time-to-event outcomes (Vandenberghe et al. In press), a remaining concern was that some patients died or experienced the event before surgery and thus before the mediator was assessed. As a result, our mediation analysis was limited to the subgroup of patients who were alive at the time of surgery, when pathological complete response was assessed. In this subgroup, the two treatment arms may no longer have been comparable and we argued that future work would thus have to include an extension of the proposed techniques to account for this. Although we found a better way to deal with this problem in **Chapter 4**, there are still some remaining questions that should be addressed in the future, which is why we provide the reader with a detailed discussion of the problem in this Section. This type of bias, also called guarantee-time bias (GTB), can occur in longitudinal analyses in which groups, that are defined via a classifying event occurring sometime during follow-up (i.e. mediator), are compared (Giobbie-Hurder et al. 2013). Giobbie-Hurder et al. (2013) describe three analytic techniques to handle immortal time bias: conditional landmark analysis, a Cox regression model with time-varying covariates, and inverse probability weighting, of which we used the former in our analysis of the EORTC 10994/BIG 1-00 trial in **Chapter 3**. We will first discuss each of the three approaches in turn, before proceeding to our own proposal from **Chapter 4** to handle this problem. First, the conditional landmark approach does not use time from randomisation till event, but time from the mediator measurement (i.e. surgery as landmark) till event. It has the huge advantage of being a very simple approach, but unlike our proposal in **Chapter 4** the disadvantage that it excludes patients who die or have an event before the landmark. Conditioning on survival up to mediator assessment could also induce bias when as a result treatment groups are no longer comparable. The second approach, a Cox regression model that includes a time-varying classifying event (i.e. mediator), overcomes these limitations by using all patient data, but does not allow long term prediction. The third approach makes use of inverse probability weights for the classifying event (i.e. mediator) given treatment, an event indicator up to that point and baseline patient characteristics and disease-related variables and fit a marginal structural model to estimate the controlled direct effect of treatment on the time-to-event outcome, in contrast to the natural direct and indirect effects that we were interested in.

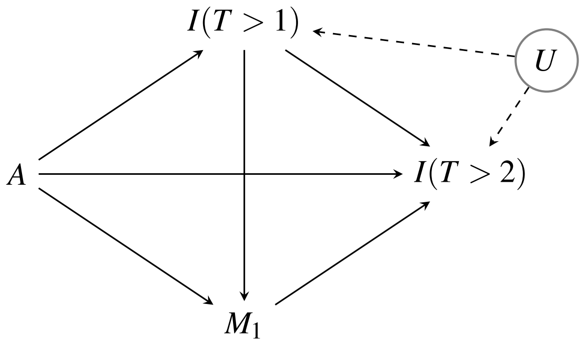


Figure 5.2: Causal diagram representation of immortal time bias.

The literature on multiple mediators that helped us with this problem would treat the event indicator $I(T > 1)$ (Figure 5.2) as an intermediate confounder (i.e. merely a common cause of the second mediator and the outcome that is affected by the exposure) and the second mediator (i.e. here M_1) as mediator of interest (Miles et al. 2017; Tchetgen Tchetgen and Shpitser 2012; Tchetgen Tchetgen and VanderWeele 2014; VanderWeele et al. 2014; Vansteelandt and VanderWeele 2013). This approach would lead to a coarse two-way decomposition of the total intention-to-treat effect into an indirect effect only via the second mediator (i.e. here M_1) and a direct effect not through this mediator. A similar strategy is used in **Chapter 4** with repeatedly measured mediators and a time-to-event outcome, where survival up to a certain time t is treated as a time-varying covariate L_t, which is 'permitted' since we allow unmeasured confounders for the relationship between time-varying counfounders and the outcome. As a result, the partial indirect effect is estimated: the effect of randomised assignment to treatment on survival as transmitted along the combination of pathways whereby treatment directly influences one of the mediators, which thus excludes pathways whereby treatment initially influences the survival indicator. As such $S_{1,0}(t)$, say for $t \in [2,3[$ would express how likely it would be to survive time t, if (a) one were assigned to the treatment and (b) provided one survived wave 1, M_1 were set to the value it would have taken if one were assigned to placebo and survived wave 1. Contrasting $S_{1,0}(t)$ with $S_{1,1}(t)$ would then express the partial indirect effect via M_1.

There are some remaining concerns about the proposal in **Chapter 4**. First, it might be that $S_{1,0}(t)$ is not very well defined when for instance the treatment has a beneficial effect on survival as compared to the placebo. Because then we would have to imagine for someone who under placebo would have died, say before *t=1*, what would have happened to their survival probability if assigned to treatment, but this person could have been kept alive ‘somehow’ under placebo, so that the value of M_1 under placebo would have been known. If treatment is indeed beneficial for all patients, it could therefore be a better solution to contrast $S_{0,0}(t)$ to $S_{0,1}(t)$ to express the indirect effect as this would avoid the need to keep patients alive ‘artificially’. Another option is to turn to alternative effect measures that are not defined in terms of cross-world counterfactuals, so-called interventional direct and indirect effects (Vansteelandt and Daniel 2017; VanderWeele and Tchetgen Tchetgen 2017; Lin et al. 2016; Zheng and van der Laan 2017). Remember that they differ from natural direct and indirect effects because they do not fix the mediator level to be equal to the counterfactual mediator value at a certain level of treatment, but to a random draw of the distribution of the mediator at a certain exposure level given covariates. As such, one would not have to imagine the patient’s mediator value under placebo if this person could have been kept alive, but we would set the mediator value to a random draw of the distribution of this mediator in survivors under placebo given covariate history (Zheng and van der Laan 2017). The simple example in **Chapter 4** where a patient died on the placebo arm due to a car crash, showed that these might not be very realistic values however and that conceptually at least natural direct and indirect effects might be more appealing. It seems worthwhile however to look for different effect measures in the future that avoid this problem all together. A final concern is the relative complexity of this method in **Chapter 4** as there is a whole process to go through before obtaining the final results. What would thus really make this research more accessible for applied researchers who are interested in these types of research questions is the development of an approach that circumvents this complexity, so that one is able to obtain effect estimates via a single model (see for example *natural effect models* (Lange et al. 2012; Vansteelandt et al. 2012a)).

CHAPTER 6

Samenvatting

De gerandomiseerde gecontroleerde studie wordt nog steeds beschouwd als de gouden standaard voor de beoordeling van een mogelijks oorzakelijk effect van een bepaalde blootstelling of behandeling (bv. chemotherapie) op een uitkomst (bv. sterfte). De zogenaamde 'intention to treat'-analyses, waarbij patiënten geanalyseerd worden volgens de behandelingsgroep waartoe ze oorspronkelijk werden gerandomiseerd, worden gebruikt om deze oorzakelijke verbanden op te sporen. Meer en meer worden deze primaire analyses van gerandomiseerde studies aangevuld met resultaten van statistische mediatie-analyses. Ondertussen is de wetenschappelijke gemeenschap er zich immers van bewust dat deze resultaten een grote meerwaarde kunnen bieden aangezien ze tot een meer diepgaand inzicht in de mogelijke processen onderliggend aan die oorzaak-gevolg relaties kunnen leiden. Aan de hand van een mediatie-analyse kan men immers nagaan in welke mate het effect van de behandeling op de uitkomst toe te schrijven is aan het effect op mogelijks tussenliggende factoren, ook wel mediatoren genoemd (bv. krimpen van de tumor), enerzijds en de aanwezigheid van andere, niet nader gedefinieerde processen anderzijds. De eerste manier waardoor de behandeling een effect kan hebben op de uitkomst via een specifieke tussenliggende factor wordt in statistische mediatie-analyse het *indirect* effect genoemd, de overige mechanismen die het behandelingseffect tot stand brengen worden gebundeld onder het *direct* effect.

In **Hoofdstuk 1** beschrijven we vier gebieden waarbij een mediatie-analyse een toegevoegde waarde kan zijn bovenop de standaard ‘intention to treat’-analyse. Als eerste leiden de resultaten van een mediatie-analyse tot een beter inzicht in hoe een bepaalde behandeling effect heeft op de uitkomst. Als we weten welke factoren er voor het gunstige behandelingseffect zorgen, dan is het eventueel mogelijk om betere en efficiëntere behandelingen met minder bijwerkingen te ontwikkelen door net deze factoren te benadrukken en minder essentiële zaken weg te laten (Marso et al. 2016). Een tweede manier waarop mediatie-analyse resultaten kunnen bijdragen is hun nut in het evalueren van de effectiviteit van bestaande behandelingen voor een probleem of ziekte waarvoor ze oorspronkelijk niet ontwikkeld werden. Uit onderzoek blijkt bijvoorbeeld dat antidepressiva niet alleen symptomen van depressie verminderen, maar ook cognitieve functies verbeteren. Deze medicijnen zouden dus niet enkel voor de behandeling van depressieve symptomen voorgeschreven kunnen worden, maar ook om bijvoorbeeld tekorten op het vlak van cognitieve functies bij ouderen te verbeteren (McIntyre et al. 2014). Als voorbeeld van een derde context waar mediatie-analyse duidelijkheid kan brengen, werd de MIRA-trial (Padian et al. 2007; Rosenblum et al. 2009) beschreven, waar het effect van de beschikbaarheid van een pessarium en glijmiddel op HIV risico niet eenduidig genterpreteerd kon worden omdat in deze behandelingsgroep significant minder vrouwen een condoom gebruikten dan in de placebo groep die geen toegang had tot een pessarium of glijmiddel. Via een mediatie-analyse kreeg Rosenblum et al. (2009) een beter zicht op het directe effect van pessarium en glijmiddel beschikbaarheid op HIV risico door het te onderscheiden van hun effect op HIV risico door hun invloed op condoomgebruik. Tenslotte kan een mediatie-analyse ook van nut zijn voor de evaluatie van mediërende factoren die indicatoren zijn van ziekte op lange termijn, zogenaamde surrogaat merkers, die gebruikt worden om klinisch relevante lange termijn uitkomsten in studies te vervangen en zo de tijd en bijhorende kosten van de studie te drukken (Vandenberghe et al. In press). De traditionele aanpak om directe en indirecte effecten te schatten is wel intuïtief en eenvoudig, maar kan beter niet gebruikt worden los van lineaire modellen. Recente ontwikkelingen binnen de causale inferentie literatuur hebben geleid tot formele definities van deze directe en indirecte effecten, die ook gebruikt kunnen worden voor non-lineaire modellen en bij interacties. Een van de belangrijkste bijdragen van deze populaire causale inferentie literatuur is dat deze directe en indirecte effecten geïdentificeerd kunnen worden aan de hand van de zogenaamde *mediation formula* (Pearl 2001, 2012) via arbitraire modellen voor de uitkomst en mediator. In de context van gerandomiseerde studies heeft deze *mediation formula* als grootste nadeel dat de resultaten vertekend kunnen zijn wanneer het model voor de mediator, het model voor de uitkomst of beide fout gespecificeerd zijn. Aangezien men het totale 'intention to treat'-effect onvertekend kan schatten in gerandomiseerde studies zonder het gebruik van statistische modellen, is men vrij sceptisch ten op zichte van analyses die door het gebruik van extra modellen wel aan efficiëntie winnen, maar tezelfdertijd ook een vertekend resultaat kunnen geven. Ook in een gerandomiseerde studie kan een mediatie-analyse echter niet uitgevoerd worden zonder enige modellering omdat de mediator (in tegenstelling tot de behandeling) niet gerandomiseerd is en men dus moet controleren voor mogelijke gemeenschappelijke oorzaken (i.e. *confounders*) van de mediator en de uitkomst. Daarom hebben wij ervoor gekozen om in **Hoofdstuk 2** en **Hoofdstuk 3** ons te focussen op meer robuste strategieën die minder gevoelig zijn voor model misspecificatie.

Een alternatieve strategie dan de *mediation formula*, die gebruik maakt van een model voor de uitkomst en de mediator, is een van deze twee modellen vervangen door een model voor de behandeling (Tchetgen Tchetgen and Shpitser 2012; Vansteelandt 2012b). Een van de contexten waarin deze strategie aangemoedigd wordt zijn gerandomiseerde studies aangezien men de randomisatie kans kent en dus zeker kan zijn dat dit model juist gespecificeerd is. In **Hoofdstuk 2** twee wordt eerst een simpele schatter besproken voor continue en binaire uitkomsten (Tchetgen Tchetgen and Shpitser 2012) waarvan de resultaten onvertekend zullen zijn als het model voor de uitkomst juist gespecificeerd wordt. Enkel het model voor de uitkomst, want het model voor de behandeling die ook gebruikt wordt is gekend en dus correct. We hadden evenzeer kunnen kiezen om het uitkomst model door dat van de behandeling te vervangen en een correcte specificatie van een model voor de mediator te eisen, maar er waren twee redenen waarom dit ons een minder gepaste aanpak leek. Als eerste zou deze methode gebruik moeten maken van invers wegen met de mediator distributie wat voor schatters kan zorgen die onstabiel gedrag vertonen, vooral bij misspecificaties in de staarten van de mediator distributie. Ten tweede is het zo dat het model voor de gemiddelde uitkomst meestal makkelijker specificeerbaar is dan een model voor de gehele mediator distributie. Wanneer er echter covariaten beschikbaar zijn die sterk voorspellend zijn voor de mediator, dan kan het niet gebruiken van dit model voor de mediator distributie leiden tot aanzienlijk minder efficiënte schattingen. Daarom stelden we in **Hoofdstuk 2** een eigen strategie voor die wel gebruik maakt van een model voor de mediator zodat alle beschikbare informatie in de covariaten benut kan worden, maar die tot onvertekende resultaten blijft leiden zelfs als dit model voor de mediator niet juist gespecificeerd is. In **Hoofdstuk 3** beschrijven we moderne mediatie-analyse technieken voor uitkomsten die tijd tot een bepaalde gebeurtenis (bv. tijd tot sterfte) betreffen en breiden we het voorstel uit **Hoofdstuk 2** uit naar dergelijke uitkomsten. Hoewel ons voorstel algemeen gebruikt kan worden voor mediatie-analyses bij *survival* uitkomsten, werd dit hoofdstuk geschreven vanuit een *surrogaat merker* vraagstelling. We tonen namelijk aan dat een mediatie-analyse informatief kan zijn voor het evalueren van surrogaat merkers wanneer men slechts over data van n enkel experiment beschikt.

Het merendeel van de recente publicaties omtrent causale mediatie-analyse is gefocust op het effect van een behandeling op de uitkomst via n specifieke mediator die op n tijdstip gemeten werd. Realistische onderzoeksvragen betreffen echter vaak meer complexe toepassingen waarbij men het behandelingseffect via meerdere mediatoren wenst te onderzoeken. In **Hoofdstuk 4** gingen we daarom op zoek naar bestaande strategieën op het gebied van longitudinale mediatie-analyse waarbij mediatoren meerdere keren gemeten worden gedurende de studieperiode. Omdat deze bestaande technieken een aantal nadelen hebben, stellen we zelf een strategie voor die gebruikt kan worden bij een gerandomiseerde behandeling, mediatoren die meerdere malen gemeten worden, tijdsafhankelijke confounders en een *survival* uitkomst. In tegenstelling tot het recente werk van Zheng and van der Laan (2017) richten we ons op *natural* direct en indirecte effecten (en niet op *interventional* direct en indirecte effecten) bij een *survival* uitkomst. Alhoewel beide voorstellen veel gemeenschappelijk hebben, zijn er toch subtiele verschillen die in detail werden besproken, maar waarover er toch nog een aantal vragen blijven die interessant kunnen zijn om verder te onderzoeken. In **Hoofdstuk 5** tenslotte eindigen we met een conclusie van de resultaten, bespreken we de beperkingen van de huidige literatuur omtrent mediatie-analyse en maken we een aantal suggesties voor toekomstig onderzoek.

Bibliography

(2013), "FDA approves Perjeta for neoadjuvant breast cancer treatment," .

(2014), "Pathological Complete Response in Neoadjuvant Treatment of High-Risk Early-Stage Breast Cancer: Use as an Endpoint to Support Accelerated Approval," .

Aalen, O. O., Cook, R. J., and Roysland, K. (2015), "Does Cox analysis of a randomized survival study yield a causal treatment effect?" *Lifetime Data Analysis*, 21, 579–593.

Albert, J. M. (2012), "Mediation Analysis for Nonlinear Models with Confounding." *Epidemiology*, 23, 879–888.

Albert, J. M. and Nelson, S. (2011), "Generalized Causal Mediation Analysis." *Biometrics*, 67, 1028–1038.

Alonso, A., Van der Elst, W., Molenberghs, G., and Buyse, M. (2015), "On the Relationship between the Causal-Inference and Meta-Analytic Paradigms for the Validation of Surrogate Endpoints," *Biometrics*, 71, 15–24.

Alvarez, E., Perez, V., Dragheim, M., Loft, H., and Artigas, F. (2012), "A double-blind, randomized, placebo-controlled, active-reference study of Lu AA21004 in patients with major depressive disorder (MDD)." *International Journal of Neuropsychopharmacology*, 15, 589–600.

Ananth, C. V. and VanderWeele, T. J. (2011), "Placental Abruption and Perinatal Mortality With Preterm Delivery as a Mediator: Disentangling Direct and Indirect Effects," *American Journal of Epidemiology*, 174, 99–108.

Avin, C., Shpitser, I., and Pearl, J. (2005), "Identifiability of Path-Specific Effects." in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, San Francisco, CA, USA: Morgan Kaufmann Publischers Inc, pp. 357–363.

Baron, R. M. and Kenny, D. A. (1986), "The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations." *Journal of Personality and Social Psychology*, 51, 1173–1182.

Bartlett, J. W., Carpenter, J. R., Tilling, K., and Vansteelandt, S. (2014), "Improving upon the efficiency of complete case analysis when covariates are mnar." *Biostatistics*, 15, 719–730.

Bentler, P. M. (1980), "Multivariate analysis with latent variables: Causal modeling." *Annual Review of Psychology*, 31, 419–456.

Bonnefoi, H., Litière, S., Piccart, M., MacGrogan, G., Fumoleau, P., Brain, E., Petit, T., Rouanet, P., Jassem, J., Moldovan, C., Bodmer, A., Zaman, K., Cufer, T., Campone, M., Luporsi, E., Malmström, P., Werutsky, G., Bogaerts, J., Bergh, J., and Cameron, D. A. (2014), "Pathological complete response after neoadjuvant chemotherapy is an independent predictive factor irrespective of breast cancer intrinsic subtypes: a landmark and two-step approach analyses from the EORTC 10994/BIG 1-00 phase III trial." *Annals of Oncology*, 25, 1128–1136.

Bonnefoi, H., Piccart, M., Bogaerts, J., Mauriac, L., Fumoleau, P., Brain, E., Petit, T., Rouanet, P., Jassem, J., Blot, E., Zaman, K., Cufer, T., Lortholary, A., Lidbrink, E., André, S., Litière, S., Dal Lago, L., Becette, V., Cameron, D. A., Bergh, J., and Iggo, R. (2011), "TP53 status for prediction of sensitivity to taxane versus non-taxane neoadjuvant chemotherapy in breast cancer (EORTC 10994/BIG 1-00): a randomised phase 3 trial." *The Lancet Oncology*, 12, 527–539.

Boulenger, J. P., Loft, H., and Olsen, C. K. (2014), "Efficacy and safety of vortioxetine (Lu AA21004), 15 and 20 mg/day: a randomized double-blind, placebo-controlled, duloxetine-referenced study in the acute treatment of adult patients with major depressive disorder." *International Journal of Clinical Psychopharmacology*, 29, 138–149.

Bränström, R., Kvillemo, P., Brandberg, Y., and Moskowitz, J. T. (2010), "Self-report Mindfulness as a Mediator of Psychological Well-being in a Stress Reduction Intervention for Cancer Patients - A Randomized Study." *Annals of Behavioral Medicine*, 39, 151–161.

Buyse, M. and Molenberghs, G. (1998), "Criteria for the validation of surrogate endpoints in randomized experiments." *Biometrics*, 54, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowksi, T., Renard, D., and H, G. (2000a), "The validation of surrogate endpoints in meta-analyses of randomized experiments." *Biostatistics*, 1, 49–67.

Buyse, M., Thirion, P., Carlson, R. W., Burzykowski, T., Molenberghs, G., and Piedbois, P. (2000b), "Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis," *The Lancet*, 356, 373–378.

Bycott, P. W. and Taylor, J. M. G. (1998), "An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker." *Controlled Clinical Trials*, 19, 555–568.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data." *Biometrika*, 96, 723–734.

Colantuoni, E. and Rosenblum, M. (2015), "Leveraging prognostic baseline variables to gain precision in randomized trials." *Statistics in Medicine*, 34, 2602–2617.

Collett, D. (2003), *Modelling Survival Data in Medical Research*, Boca Raton, Fla.: Chapman and Hall/CRC.

Contiero, P., Berrino, F., Tagliabue, G., Mastroianni, A., Di Mauro, M. G., Fabiano, S., Annulli, M., and Muti, P. (2013), “Fasting blood glucose and long-term prognosis of non-metastatic breast cancer: a cohort study.” *Breast Cancer Research and Treatment*, 138, 951–959.

Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J. P., Wolmark, N., Bonnefoi, H., Cameron, D., Gianni, L., Valagussa, P., Swain, S. M., Prowell, T., Loibl, S., Wickerham, D. L., Bogaerts, J., Baselga, J., Perou, C., Blumenthal, G., Blohmer, J., Mamounas, E. P., Bergh, J., Semiglazov, V., Justice, R., Eidtmann, H., Paik, S., Piccart, M., Sridhara, R., Fasching, P. A., Slaets, L., Tang, S., Gerber, B., Geyer Jr, C. E., Pazdur, R., Ditsch, N., Rastogi, P., Eiermann, W., and von Minckwitz, G. (2014), “Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis.” *The Lancet*, 9938, 164–172.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015), “Causal Mediation Analysis with Multiple Mediators.” *Biometrics*, 71, 1–14.

Daniels, M. J. and Hughes, M. D. (1997), “Meta-analysis for the evaluation of potential surrogate markers.” *Statistics In Medicine*, 16, 1965–1982.

De Ruddere, L., Goubert, L., Vervoort, T., Kappesser, J., and Crombez, G. (2013), “Impact of being primed with social deception upon observer responses to others’ pain.” *Pain*, 154, 221–226.

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., and Micali, N. (2015), “Mediation Analysis With Intermediate Confounding: Structural Equation Modeling Viewed Through the Causal Inference Lens.” *American Journal of Epidemiology*, 181, 64–80.

Duncan, O. D. (1966), “Path analysis: Sociological examples.” *American Journal of Sociology*, 72, 1–16.

Emsley, R., Dunn, G., and White, I. A. (2010), “Mediation and moderation of treatment effects in randomised controlled trials of complex interventions.” *Statistical Methods in Medical Research*, 19, 237–270.

Fosen, J., Ferkingstad, E., Borgan, O., and Aalen, O. O. (2006), "Dynamic path analysis - a new approach to analyzing time-dependent covariates." *Lifetime Data Analysis*, 12, 143–167.

Freedman, L. S. (2001), "Confidence intervals and statistical power of the "validation" ratio for surrogate or intermediate endpoints." *Journal of Statistical Planning and Inference*, 96, 143–153.

Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992), "Statistical validation of intermediate endpoints for chronic deseases," *Statistics in Medicine*, 11.

Gail, M. H., Pfeiffer, R., van Houwelingen, H. C., and Carroll, R. J. (2000), "On meta-analytic assessment of surrogate outcomes." *Biostatistics*, 1, 231–246.

Giobbie-Hurder, A., Gelber, R. D., and Regan, M. M. (2013), "Challenges of Guarantee-Time Bias." *Journal of Clinical Oncology*, 31, 2963–2969.

Glass, T. A., Goodman, S. N., Hernan, M. A., and Samet, J. M. (2013), "Causal inference in public health." *Annual Review of Public Health*, 34, 61–75.

Goss, P. E., Smith, I. E., O'Shaughnessy, J., Ejlertsen, B., Kaufmann, M., Boyle, F., Buzdar, A. U., Fumoleau, P., Gradishar, W., Martin, M., Moy, B., Piccart-Gebhart, M., Pritchard, K. I., Lindquist, D., Chavarri-Guerra, Y., Aktan, G., Rappold, E., Williams, L. S., and Finkelstein, D. M. (2013), "Adjuvant lapatinib for women with early-stage HER2-positive breast cancer: a randomised, controlled, phase 3 trial," *The Lancet Oncology*, 14, 88–96.

Greenland, S. and Finkle, W. D. (1995), "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses," *American Journal of Epidemiology*, 142, 1255–1264.

Greenland, S., Robins, J. M., and Pearl, J. (1999), "Confounding and Collapsibility in Causal Inference." *Statistical Science*, 14, 29–46.

Group, M. R. F. I. T. R. (1990), "Mortality rates after 10.5 years for participants in the Multiple Risk Factor Intervention Trial." *Journal of the American Medical Association*, 263, 1795–1801.

Hafeman, D. M. and Schwartz, S. (2009), “Opening the black box: a motivation for the assessment of mediation,” *International Journal of Epidemiology*, 38, 838–845.

Hernan, M. A. (2005), “Invited commentary: Hypothetical interventions to define causal effects - afterthought or prerequisite?” *American Journal of Epidemiology*, 162, 618–620.

— (2010), “The Hazards of Hazard Ratios,” *Epidemiology*, 21, 13–15.

Hernan, M. A. and VanderWeele, T. J. (2011), “Compound treatments and transportability of causal inference.” *Epidemiology*, 22, 368–377.

Huang, Y.-T. and Cai, T. (2016), “Mediation Analysis for Survival Data Using Semiparametric Probit Models,” *Biometrics*, 72, 563–574.

Huang, Y.-T. and Yang, H.-I. (2017), “Causal Mediation Analysis of Survival Outcome with Multiple Mediators.” *Epidemiology*, 28, 370–378.

Imai, K., Keele, L., and Tingley, D. (2010), “A general approach to causal mediation analysis,” *Psychological Methods*, 15, 309–334.

Imai, K. and Yamamoto, T. (2013), “Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments.” *Political Analysis*, 21, 141–171.

Investigators, C. A. S. T. (1989), “Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction.” *New England Journal of Medicine*, 321, 406–412.

Joffe, M. M., Byrne, C., and Colditz, G. A. (2001), “Postmenopausal Hormone Use, Screening, and Breast Cancer: Characterization and Control of a Bias.” *Epidemiology*, 12, 429–438.

Joffe, M. M. and Greene, T. (2009), “Related Causal Frameworks for Surrogate Outcomes,” *Biometrics*, 65, 530–538.

Judd, C. M. and Kenny, D. A. (1981), “Process analysis: Estimating mediation in treatment evaluations,” *Evaluation Review*, 5, 602–619.

Kaufman, J. S. (2009), “Commentary: Gliding the black box.” *International Journal of Epidemiology*, 38, 845–847.

Kaufman, J. S., MacLehose, R. F., and Kaufman, S. (2004), “A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation,” *Epidemiologic Perspectives and Innovations*, 1.

Kraemer, H. C., Wilson, G. T., Fairburn, G. C., and Agras, W. S. (2002), “Mediators and Moderators of Treatment Effects in Randomized Clinical Trials.” *Archives of General Psychiatry*, 59, 877.

Lange, T. and Hansen, J. V. (2011), “Direct and Indirect Effects in a Survival Context,” *Epidemiology*, 22, 575–581.

Lange, T., Rasmussen, M., and Thygesen, L. C. (2014), “Assessing natural direct and indirect effects through multiple pathways.” *American Journal of Epidemiology*, 179, 513–518.

Lange, T., Vansteelandt, S., and Bekaert, M. (2012), “A Simple Unified Approach for Estimating Natural Direct and Indirect Effects,” *American Journal of Epidemiology*, 176, 190–195.

Lendle, S. D., Subbaraman, M. S., and van der Laan, M. J. (2013), “Identification and efficient estimation of the natural direct effect among the untreated.” *Biometrics*, 69, 310–317.

Liedtke, C., Mazouni, C., Hess, K. R., André, F., Tordai, A., Mejia, J. A., Symmans, W. F., Gonzalez-Angulo, A. M., Hennessy, B., Green, M., Cristofanilli, M., Hortobagyi, G. N., and Pusztai, L. (2008), “Response to Neoadjuvant Therapy and Long-Term Survival in Patients With Triple-Negative Breast Cancer,” *Journal of Clinical Oncology*, 26, 1275–1281.

Lin, D. Y., Fleming, T. R., and DeGruttola, V. (1997), “Estimating the proportion of treatment effect explained by a surrogate marker,” *Statistics in Medicine*, 16, 1515–1527.

Lin, S.-H., Young, J. G., Logan, R., and VanderWeele, T. J. (2016), "Mediation analysis for a survival outcome with time-varying exposures, mediators and confounders." *Harvard University Biostatistics Working Paper Series*, Paper 203.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., Steen, J., and Vansteelandt, S. (2013), "Flexible Mediation Analysis in the Presence of Nonlinear Relations: Beyond the Mediation Formula." *Multivariate Behavioral Research*, 48, 871–894.

Lynch, K. G., Cary, M., Gallop, R., and Ten Have, T. R. (2008), "Causal mediation analyses for randomized trials," *Health Services and Outcomes Research Methods*, 8, 57–76.

MacKinnon, D. P. (2008), *Introduction to Statistical Mediation Analysis*, Multivariate Applications Series, Lawrence Erlbaum Associates.

MacKinnon, D. P., Warsi, G., and Dwyer, H. (1995), "A simulation study of mediated effect measures." *Multivariate Behavioral Research*, 30, 41–62.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J. F. E., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., Ravn, L. S., Steinberg, W. M., Stockner, M., Zinman, B., Bergenstal, R. M., and Buse, J. B. (2016), "Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes." *The New England Journal of Medicine*, 375, 311–322.

Martinussen, T. and Vansteelandt, S. (2013), "On collapsibility and confounding bias in Cox and Aalen regression models." *Lifetime Data Analysis*, 19, 279–296.

McIntyre, R. S., Cha, D. S., Soczynska, J. K., Woldeyohannes, H. O., Gallaugher, L. A., Kudlow, P., Alsuwaidan, M., and Baskaran, A. (2013), "Cognitive deficits and functional outcomes in major depressive disorder: determinants, substrates, and treatment interventions." *Depression and Anxiety*, 30, 515–527.

McIntyre, R. S., Lophaven, S., and Olsen, C. K. (2014), "A randomized, double-blind, placebo-controlled study of vortioxetine on cognitive function in depressed adults." *International Journal of Neuropsychopharmacology*, 17, 1557–1567.

Michiels, S., Le Maître, A., Buyse, M., Burzykowski, T., Maillard, E., Bogaerts, J., Vermorken, J. B., Budach, W., Pajak, T. F., Ang, K. K., Bourhis, J., Pignon, J. P. o. b. o. t. M., and Groups, M.-N. C. (2009), "Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta- analyses of individual patient data." *Lancet Oncology*, 10, 341–350.

Mieog, J. S., van der Hage, J. A., and van de Velde, C. J. (2007), "Preoperative chemotherapy for women with operable breast cancer." *Cochrane Database of Systematic Reviews*, 2.

Miles, C. H., Shpitser, I., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2017), "Quantifying an Adherence Path-Specific Effect of Antiretroviral Therapy in the Nigeria PEPFAR Program," *Journal of the American Statistical Association*.

Moore, K. L. and van der Laan, M. J. (2009), "Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation." *Statistics in Medicine*, 28, 39–64.

Naimi, A. I. (2015), "Invited Commentary: Boundless Science - Putting Natural Direct and Indirect Effects in a Clearer Empirical Context." *American Journal of Epidemiology*, 182, 109–114.

Naimi, A. I., Kaufman, J. S., and MacLehose, R. F. (2014), "Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects." *International Journal of Epidemiology*, 43, 1656–1661.

Nandi, A., Glymour, M., Kawachi, I., and VanderWeele, T. J. (2012), "Using Marginal Structural Models to Estimate the Direct Effect of Adverse Childhood Social Conditions on Onset of Heart Disease, Diabetes, and Stroke." *Epidemiology*, 22, 223–232.

Oakley, A., Strange, V., Bonell, C., Allen, E., and Stephenson, J. (2006), "Process evaluation in randomised controlled trials of complex interventions." *British Medical Journal*, 332, 413–416.

Oba, K., Paoletti, X., Alberts, S., Bang, Y. J., Benedetti, J., Bleiberg, H., Catalano, P., Lordick, F., Michiels, S., Morita, S., Ohashi, Y., Pignon, J. P., Rougier, P., Sasako, M., Sakamoto, J., Sargent, D., Shitara, K., Van Cutsem, E., Buyse, M., and Burzykowski, T. o. b. o. t. G. g. (2013), "Disease-free survival as a surrogate for overall survival in adjuvant trials of gastric cancer: a meta-analysis." *Journal of the National Cancer Institute*, 5, 1600–1607.

Oba, K., Sato, T., Ogihara, T., Satura, T., and Nakao, K. (2011), "How to use marginal structural models in randomized trials to estimate the natural direct and indirect effects of therapies mediated by causal intermediates." *Clinical Trials*, 8, 277–287.

Padian, N. S., van der Straten, A., Ramjee, G., Chipato, T., de Bruyn, G., Blanchard, K., Shiboski, S., Montgomery, E. T., Fancher, H., Cheng, H., Rosenblum, M., van der Laan, M. J., Jewell, N., and McIntyre, J. (2007), "Diaphragm and lubricant gel for prevention of HIV acquisition in southern African women: a randomised controlled trial." *The Lancet*, 370, 251–261.

Pearl, J. (2001), "Direct and indirect effects." Morgan Kaufmann: San Francisco, pp. 411–420.

— (2012), "The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models." in *Causality: Statistical Perspectives and Applications*, eds. Berzuini, C. L. B., Dawid, P., and Bernardinelli, L. e. a., Chichester, UK: John Wiley and Sons, pp. 151–179.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006), "Estimation of direct causal effects." *Epidemiology*, 17, 276–284.

Pirl, W. F., Greer, J. A., Traeger, L., Jackson, V., Lennes, I. T., Gallagher, E. R., Perez-Cruz, P., Heist, R. S., and Temel, J. S. (2012), "Depression and Survival in Metastatic Non–Small-Cell Lung Cancer: Effects of Early Palliative Care," *Journal of Clinical Oncology*, 30, 1310–1315.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002), "Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems." *Statistics in Medicine*, 21, 2917–2930.

Preacher, K. J. and Hayes, A. F. (2008), “Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models.” *Behavior Research Methods*, 40, 879–891.

Prentice, R. L. (1989), “Surrogate endpoints in clinical trials: Definition and operational criteria.” *Statistics in Medicine*, 8, 431–440.

Rasmussen, S. and Nordisk, N. (2016), “LEADER: Liraglutide and cardiovascular outcomes in type 2 diabetes,” .

Robins, J. M. (2003), “Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects.” in *Highly Structured Stochastic Systems*, eds. Green, P., Hjort, N., and Richardson, S., Oxford University Press, New York, pp. 70–81.

Robins, J. M. and Greenland, S. (1992), “Identifiability and Exchageability for Direct and Indirect Effects.” *Epidemiology*, 3, 143–155.

Robins, J. M. and Richardson, T. S. (2011), *Causality and Psychopathology.*, Oxford University Press, Oxford.

Rochon, J., du Bois, A., and Lange, T. (2014), “Mediation analysis of the relationship between institutional research activity and patient survival.” *BMC Medical Research Methodology*, 14.

Rosenblum, M., Jewell, N. P., van der Laan, M. J., Shiboski, S., van der Straten, A., and Padian, N. (2009), “Analysing direct effects in randomized trials with secondary interventions: an application to human immunodeficiency virus prevention trials.” *Journal of the Royal Statistical Society A*, 172, 443–465.

Rosenblum, M. and van der Laan, M. J. (2009), “Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models,” *Biometrics*, 65, 937–945.

Rubin, D. B. and van der Laan, M. J. (2008), “Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis.” *International Journal of Biostatistics*, 4, Article 5.

Sargent, D. J., Wieand, H. S., Haller, D. G., Gray, R., Benedetti, J., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., Francini, G., Grothey, A., O'Connell, M., Catalano, P. J., Blanke, C. D., Kerr, D., Green, E., Wolmark, N., Andre, T., Goldberg, R. M., and de Gramont, A. (2005), "Disease- free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials." *Journal of Clinical Oncology*, 23.

Schwartz, S., Hafeman, D. M., Campbell, U., and Gatto, N. (2010), "Authors' repsonse to: Commentary: Gilding the black box." *International Journal of Epidemiology*, 39.

Senn, S. (2000), "Consensus and Controversy in Pharmaceutical Statistics." *Journal of the Royal Statistical Society D*, 49, 135–176.

Simon, H. A. (1954), "Spurious correlation: A causal interpretation." *Journal of the American Statistical Association*, 49, 467–479.

Steen, J., Loeys, T., Moerkerke, B., and Vansteelandt, S. (2016), "Medflex: An R Package for Flexible Mediation Analysis Using Natural Effect Models." *Journal of Statistical Software*, in press.

— (2017), "Flexible mediation analysis with multiple mediators." *American Journal of Epidemiology*, in press.

Strohmaier, S., Roysland, K., Hoff, R., Borgan, O., Pedersen, T. R., and Aalen, O. O. (2015), "Dynamic path analysis - a useful tool to investigate mediation processes in clinical survival trials." *Statistics in Medicine*, 34, 3866–3887.

Taguri, M., Featherstone, J., and Cheng, J. (2015), "Causal mediation analysis with multiple causally non-ordered mediators." *Statistical Methods in Medical Research*, In press.

Taylor, A. B., MacKinnon, D. P., and Tein, J.-Y. (2008), "Tests of the Three-Path Mediated Effect," *Organizational Research Methods*, 11, 241–269.

Tchetgen Tchetgen, E. J. (2011), "On causal mediation analysis with a survival outcome," *The international journal of biostatistics*, 7, Article 33.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012), "Semiparametric Theory for Causal Mediation Analysis: efficiency bounds, multiple robustness, and sensitivity analysis." *Annals of Statistics*, 40, 1816–1845.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2014), "On identification of natural direct effects when a confounder of the mediator is directly affected by exposure." *Epidemiology*, 25, 282–291.

Tein, J.-Y. and MacKinnon, D. P. (2003), "Estimating Mediated Effects with Survival Data." in *New Developments in Psychometrics.*, eds. Yanai, H., Okada, A., Shigemasu, K., Kano, Y., and Meulman, J. J., Tokyo: Springer.

Ten Have, T. R. and Joffe, M. M. (2012), "A review of causal estimation of effects in mediation analyses." *Statistical Methods in Medical Research*, 21, 77–107.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008), "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach." *Statistics in Medicine*, 27, 4658–4677.

Valeri, L. and VanderWeele, T. J. (2013), "Mediation Analysis Allowing for Exposure–Mediator Interactions and Causal Interpretation: Theoretical Assumptions and Implementation With SAS and SPSS Macros." *Psychological Methods*, 18, 137–150.

van der Laan, M. J. and Petersen, M. L. (2008), "Direct Effect Moddels." *The international journal of biostatistics*, 4, Article 23.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), "Super Learner." *Statistical Applications in Genetics and Molecular Biology*, 6, 1–21.

Vandenberghe, S., Duchateau, L., Slaets, L., Bogaerts, J., and Vansteelandt, S. (In press), "Surrogate marker analysis in cancer clinical trials through time-to-event mediation techniques." *Statistical Methods in Medical Research*.

Vandenberghe, S., Vansteelandt, S., and Loeys, T. (2017a), "Boosting the precision of mediation analyses of randomised experiments through covariate adjustment," *Statistics In Medicine*, 36, 939–957.

VanderWeele, T. J. (2009), "Marginal Structural Models for the Estimation of Direct Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

— (2011), "Causal mediation analysis with survival data." *Epidemiology*, 22, 582–585.

— (2013), "Surrogate Measures and Consistent Surrogates." *Biometrics*, 69, 561–581.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017), "Mediation analysis with time varying exposures and mediators," *Journal of the Royal Statistical Society B*, 79, 917–938.

VanderWeele, T. J. and Vansteelandt, S. (2009), "Conceptual issues concerning mediation, interventions and composition." *Statistics and its Interface*, 2, 457–468.

— (2010), "Odds Ratios for Mediation Analysis for a Dichotomous Outcome," *American Journal of Epidemiology*, 172, 1339–13348.

— (2013), "Mediation Analysis with Multiple Mediators." *Epidemiologic Methods*, 2, 95–115.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014), "Effect decomposition in the presence of an exposure-induced mediator-outcome confounder." *Epidemiology*, 25, 300–306.

Vansteelandt, S. (2012b), "Understanding Counterfactual-Based Mediation Analysis Approaches and Their Differences." *Epidemiology*, 23, 889–891.

Vansteelandt, S., Bekaert, M., and Lange, T. (2012a), "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects." *Epidemiologic Methods*, 1, Article 7.

Vansteelandt, S. and Daniel, R. M. (2017), "Interventional Effects for Mediation Analysis with Multiple Mediators." *Epidemiology*, 28, 258–265.

Vansteelandt, S. and VanderWeele, T. J. (2013), "Natural Direct and Indirect Effects on the Exposed: Effect Decomposition under Weaker Assumptions." *Biometrics*, 68, 1019–1027.

Vermeulen, K., Thas, O., and Vansteelandt, S. (2015), "Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment." *Statistics in Medicine*, 34, 1012–1030.

Wang, J., Spitz, M. R., Amos, C. I., Wilkinson, A. V., Wu, X., and Shete, S. (2010), "Mediating Effects of Smoking and Chronic Obstructive Pulmonary Disease on the Relation Between the CHRNA5-A3 Genetic Locus an Lung Cancer Risk." *Cancer*, 116, 3458–3462.

Wright, S. (1920), "The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs." *Proceedings of the National Academy of Sciences*, 6, 320–332.

Yang, L. and Tsiatis, A. A. (2001), "Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial." *The American Statistician*, 55, 314–321.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008), "Improving efficiency of inferences in randomized clinical trials using auxiliary covariates." *Biometrics*, 64, 707–715.

Zheng, W. and van der Laan, M. J. (2012a), "Targeted Maximum Likelihood Estimation of Natural Direct Effects." *The international journal of biostatistics*, 8, Article 3.

— (2012b), "Causal Mediation in a Survival Setting with Time-Dependent Mediators." Technical Report 295, Division of Biostatistics, University of California, Berkeley.

— (2017), "Longitudinal Mediation Analysis with Time-varying Mediators and Exposures, with Application to Survival Outcomes." *Journal of Causal Inference*, In press.