Medflex: an R package for flexible mediation analysis using natural effect models

Johan Steen Ghent University
Tom Loeys Ghent University
Beatrijs Moerkerke Ghent University
Stijn Vansteelandt Ghent University

Abstract

Mediation analysis is routinely adopted by researchers from a wide range of applied disciplines as a statistical tool to disentangle the causal pathways by which an exposure or treatment affects an outcome. The counterfactual framework provides a language for clearly defining path-specific effects of interest and has fostered a principled extension of mediation analysis beyond the context of linear models. This paper describes **medflex**, an R package that implements some recent developments in mediation analysis embedded within the counterfactual framework. The **medflex** package offers a set of ready-made functions for fitting natural effect models, a novel class of causal models which directly parameterize the path-specific effects of interest, thereby adding flexibility to existing software packages for mediation analysis, in particular with respect to hypothesis testing and parsimony. In this paper, we give a comprehensive overview of the functionalities of the **medflex** package.

Keywords: causal inference, mediation analysis, direct effect, indirect effect, natural effect models, **medflex**, R.

1. Introduction

Empirical studies often aim at gaining insight into the underlying mechanisms by which an exposure or treatment affects an outcome of interest. Mediation analysis, as popularized in psychology and the social sciences by Judd and Kenny (1981) and Baron and Kenny (1986), has been widely adopted as a statistical tool to shed light on these mechanisms, by enabling the decomposition of total causal effects into an *indirect* effect through a hypothesized intermediate variable or mediator and the remaining *direct* effect. Although its initial formulations were restricted to the context of linear regression models, several attempts have been made to extend the application of traditional estimators for indirect effects (i.e., product-of-coefficients and difference-in-coefficients estimators) beyond linear settings (e.g., MacKinnon and Dwyer 1993; MacKinnon, Lockwood, Brown, Wang, and Hoffman 2007; Hayes and Preacher 2010; Iacobucci 2012). However, these extensions lack formal justification and yield effect estimates that are often difficult to interpret (e.g., Pearl 2012).

Recent advances from the causal inference literature (e.g., Albert 2008; Albert and Nelson 2011; Avin, Shpitser, and Pearl 2005; Imai, Keele, and Yamamoto 2010b; Pearl 2001, 2012; Robins and Greenland 1992; VanderWeele and Vansteelandt 2009, 2010) have furthered these earlier attempts and improved both inference and interpretability of causal effect estimators in nonlinear settings by building on the central notion of counterfactual or potential out-

comes. This notion provides a framework that has aided in (i) formally defining direct and indirect effects (in a way that is not tied to a specific statistical model), (ii) describing the conditions required for their identification (unveiling and formalizing often implicitly made causal assumptions) and (iii) assessing the robustness of empirical findings against violations of these identification conditions (i.e., sensitivity analysis).

For instance, Imai, Keele, and Tingley (2010a) proposed mediation analysis techniques that can be applied within a larger class of nonlinear models. They implemented these in a user-friendly R package, called **mediation** (Tingley, Yamamoto, Hirose, Keele, and Imai 2014; see Hicks and Tingley 2011 for a version in Stata (StataCorp 2013) with more limited functionality). More recently, Valeri and VanderWeele (2013) reviewed the latest developments in mediation analysis for nonlinear models, focusing on exposure-mediator interactions, and provided SAS (SAS Institute Inc. 2014) and SPSS (IBM Corporation 2013) macros, enabling practitioners to easily conduct these methods using well-known commercial packages. Similarly, Emsley and Liu (2013) and Muthén and Asparouhov (2015) described how direct and indirect effects as defined in the counterfactual framework can be estimated in Stata and via extended types of structural equation models in Mplus (Muthén and Muthén 1998-2012), respectively.

In this paper, we introduce **medflex**, an R package that allows for flexible estimation of direct and indirect effects while accommodating some of the limitations of other available packages. More specifically, we make use of novel so-called *natural effect models* (Lange, Vansteelandt, and Bekaert 2012; Lange, Rasmussen, and Thygesen 2014; Loeys, Moerkerke, De Smet, Buysse, Steen, and Vansteelandt 2013; Vansteelandt, Bekaert, and Lange 2012), which directly parameterize the target causal estimands on their most natural scale. This renders formal testing and interpretation more straightforward compared to other approaches as implemented in the aforementioned software applications. The **medflex** package is freely available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=medflex` (R Core Team 2014).

Throughout, the functionalities of the **medflex** package will be illustrated using data from a survey study that was part of the Interdisciplinary Project for the Optimization of Separation trajectories (IPOS). This large-scale project involved the recruitment of individuals who divorced between March 2008 and March 2009 in four major courts in Flanders. It aimed to improve the quality of life in families during and after the divorce by translating research findings into practical guidelines for separation specialists (such as lawyers, judges, psychologists, welfare workers...) and by promoting evidence-based policy. The corresponding dataset (`UPBdata`) is included in the package and involves a subsample of 385 individuals who responded to a battery of questionnaires related to romantic *relationship* characteristics (such as adult attachment style) and *breakup* characteristics (such as breakup initiator status, experiencing negative affectivity and engaging in unwanted pursuit behaviors; UPB) (De Smet, Loeys, and Buysse 2012). Respondents were asked to imagine their former partner as well as possible and to remember how they generally felt in their relationship *before* the breakup when completing the attachment style questionnaire. The mediation hypothesis of interest concerned the question whether the level of emotional distress or negative affectivity experienced *during* the breakup can be regarded as an intermediate mechanism (M) through which attachment style towards the ex-partner *before* the breakup (X) exerts its influence on displaying UPBs *after* the breakup (Y) (Loeys *et al.* 2013).

In the next section, we briefly introduce the mediation formula (Pearl 2001, 2012), which

is the predominant vehicle for effect decomposition within the counterfactual framework. Advantages of natural effect models over direct application of the mediation formula will also be discussed in more detail. We then explain how to fit natural effect models, focusing on two missing data techniques for fitting these models and demonstrate how these approaches can be implemented in R using the **medflex** package (section 3). Next, we demonstrate how different types of exposure and mediator variables can be dealt with (section 4) and how to assess effect modification of natural effects (e.g., exposure-mediator interactions and moderated mediation) (section 5). Tools are provided for deriving and visualizing different causal effects estimates (section 6) and for estimating population-average natural effects (section 7) and natural indirect effects as defined through multiple intermediate pathways jointly (section 8). Finally, in section 9, we give some concluding remarks and list some extensions of the package which are planned to be implemented in the near future.

2. The mediation formula

2.1. Counterfactual outcomes and effect decomposition

A major appeal of the counterfactual framework is that it enables to decompose the total causal effect into a so-called *natural* direct and *natural* indirect effect, irrespective of the data distribution or scale of the effect (Robins and Greenland 1992).

Suppose that the outcome Y of an individual i that would have been observed if, possibly contrary to the fact, that individual would have been assigned to treatment arm x (or would have been exposed at exposure level x), is represented by the counterfactual or potential outcome $Y_i(x)$. The total causal effect of X on Y (for individual i) corresponding to a one-unit change in exposure or treatment level can then be derived by comparing $Y_i(1)$ and $Y_i(0)$. For instance, on an additive scale, the individual total causal effect can be expressed as $Y_i(1)-Y_i(0)$. Adopting this counterfactual notation naturally leads to framing causal inference as a missing data problem: for each individual i, only $Y_i(X_i)$ $(= Y_i)$ is observed. This missing data problem has been referred to by Holland (1986) as the 'Fundamental Problem of Causal Inference'. As a result, average causal effects $\mathsf{E}(Y(1) - Y(0))$ can only be estimated by $\mathsf{E}(Y|X = 1) - \mathsf{E}(Y|X = 0)$ under the assumption that there is no confounding between X and Y. This ignorability assumption is often expressed in terms of the following counterfactual independence:

$$Y(x) \perp\!\!\!\perp X.$$

In observational studies, this assumption is usually unrealistic: as exposed subjects typically differ from unexposed subjects, the average causal effect $\mathsf{E}(Y(1) - Y(0))$ can no longer be estimated by $\mathsf{E}(Y|X = 1) - \mathsf{E}(Y|X = 0)$ without bias, as in randomized experiments. Such violations are typically remedied by instead assuming *conditional* ignorability or (conditional) independence within strata of a given set of measured baseline covariates C:

$$Y(x) \perp\!\!\!\perp X|C,$$

so that conditional causal effects $\mathsf{E}(Y(1) - Y(0)|C)$ can still be estimated by $\mathsf{E}(Y|X = 1, C) - \mathsf{E}(Y|X = 0, C)$ without bias. This weaker assumption, however, implies that a given set of measured baseline covariates C is deemed sufficient to control for confounding, and is therefore often referred to as the assumption of 'no unmeasured confounding' (Robins 1992).

The traditional notion of direct effects corresponds to that of so-called *controlled* direct effects, which, e.g., on the additive scale, express the expected change in Y induced by a one-unit change in X while controlling the mediator M uniformly at a fixed level m for all subjects:

$$\mathsf{E}\{Y(1,m) - Y(0,m)\}.$$

Counterfactual outcomes $Y(1,m)$ and $Y(0,m)$ correspond to values the outcome would have taken if the mediator were set to a fixed level m, while, at the same time, exposure levels were set to either 1 or 0. As indicated by Robins and Greenland (1992) and Pearl (2001), this notion does not provide an equivalent operational definition of an indirect effect since it is impossible to control any of the variables in such a way that the effect of X on Y circumvents the direct pathway. Robins and Greenland (1992) introduced an alternative definition to overcome this limitation. Considering mediator levels $M(x)$ that would naturally have been observed under exposure level x, rather than a fixed mediator level m, leads to a definition of the direct effect that allows for natural variation in mediator levels and also provides a complementary definition for the indirect effect. This alternative definition calls for the introduction of *nested* counterfactuals $Y(x, M(x^*))$, which play a key role for effect decomposition in the counterfactual framework. The natural direct effect

$$\mathsf{E}\{Y(1,M(0)) - Y(0,M(0))\}$$

then expresses the expected change in Y induced by a one-unit change in X while keeping M fixed at mediator levels that would naturally have been observed if X was left unchanged (at 0). Similarly, the natural indirect effect

$$\mathsf{E}\{Y(1,M(1)) - Y(1,M(0))\}$$

expresses the expected change in Y induced by altering the levels of M from those that would naturally have been observed if X were left unchanged (at 1) to those that we would obtain if X were set to 0, while simultaneously keeping X fixed at its original value. It can easily be seen that these two expressions add to the expected total effect $\mathsf{E}\{Y(1) - Y(0)\}$ under the composition assumption that $Y(x, M(x)) = Y(x)$ (VanderWeele and Vansteelandt 2009).

As for total causal effects, identification of natural direct and indirect effects relies on strong structural assumptions. In the context of mediation analysis, the identification assumptions can be encoded in a causal diagram interpreted as a non-parametric structural equation model with independent errors. Under such diagram, identification is possible under a set of conditional independencies reflecting that a given set of baseline covariates C is sufficient to control not only for (A1) confounding between X and Y, but also for (A2) confounding between X and M, and (A3) between M and Y (after adjustment for X), and that (A4) no confounders of the M-Y relationship are affected by X (i.e., no exposure-induced confounding) (Imai *et al.* 2010b; Pearl 2001; VanderWeele and Vansteelandt 2009). Whereas the first two assumptions by definition hold in randomized experiments, the latter two do not. Although Judd and Kenny (1981) initially pointed to its importance, assumption (A3) since has largely been ignored in much of the social sciences literature, as witnessed by many mediation studies not controlling for confounders of the M-Y relationship. In recent years, however, this issue has been brought back to attention within the social sciences (e.g., Bullock, Green, and Ha 2010; MacKinnon 2008).

2.2. The mediation formula

The language of counterfactuals has enabled to clearly define causal effects in a more generic, non-parametric way, but has also promoted a more principled approach to estimating these effects than the one offered by the traditional SEM literature from the social sciences, which was mainly entrenched in linear analysis. For this purpose, the mediation formula (Pearl 2001, 2012) plays a pivotal role. It prescribes estimating the expected value of nested counterfactuals by standardizing predictions from the outcome model corresponding to exposure level x under the mediator distribution corresponding to exposure level x^*:

$$\mathsf{E}\left\{Y(x, M(x^*))|C\right\} = \sum_m \mathsf{E}(Y|X = x, M = m, C) \Pr(M = m|X = x^*, C).$$

This weighted sum can be calculated based on any type of statistical model and has been shown to yield closed-form expressions for the natural indirect effect that encompass the traditional difference-in-coefficients and product-of-coefficient estimators when confined to strictly linear models (e.g., VanderWeele and Vansteelandt 2009; Pearl 2012). However, as soon as moving beyond linear settings, the latter estimators no longer coincide with their corresponding mediation formula expressions and no longer yield readily interpretable causal effect estimates (as formalized in the counterfactual framework).[1]

More recently, closed-form expressions for natural direct and indirect effects as defined on both additive and ratio scales have been derived for a limited number of nonlinear scenarios (VanderWeele and Vansteelandt 2009, 2010; Valeri and VanderWeele 2013).

2.3. Applying the mediation formula in practice

Software applications for obtaining closed-form solutions derived from the mediation formula, as well as their corresponding delta method (or bootstrap) standard errors, have been made available as SPSS and SAS macros (Valeri and VanderWeele 2013) and as the Stata module **PARAMED** (Emsley and Liu 2013). More recently, Muthén and Asparouhov (2015) demonstrated how natural effect estimates can be obtained via extended types of structural equation models in Mplus, even in the presence of latent variables. However, such closed-form expressions can often not readily be obtained, for instance when combining a linear model for the mediator and a logistic model for the outcome.

Imai *et al.* (2010b) addressed this issue and instead suggested a more generic approach based on Monte-Carlo integration methods, which they implemented in the R package **mediation** (Tingley *et al.* 2014). Whereas its lightweight version in Stata (Hicks and Tingley 2011) and the Stata module **gformula** (Daniel, De Stavola, and Cousens 2011), which adopts a similar simulation-based approach, are restricted to parametric models, this R package also allows to specify semi- or non-parametric models for the mediator and outcome. Despite being computationally intensive, these offer more flexibility than the applications based on a purely analytical approach. In addition, the **mediation** package offers useful extensions, such as methods for dealing with multiple mediators and treatment noncompliance, while at the same time enabling users to evaluate the robustness of their findings to potential unmeasured confounding in a widely applicable sensitivity analysis.

[1] Muthén and Asparouhov (2015) give an intuitive account for SEM practitioners explaining why the product-of-coefficient estimator fails when applied in nonlinear settings or settings involving exposure-mediator interactions. Nonetheless, the product-of-coefficients method can still be useful for testing the null hypothesis of no indirect effect (VanderWeele 2011; Vansteelandt *et al.* 2012).

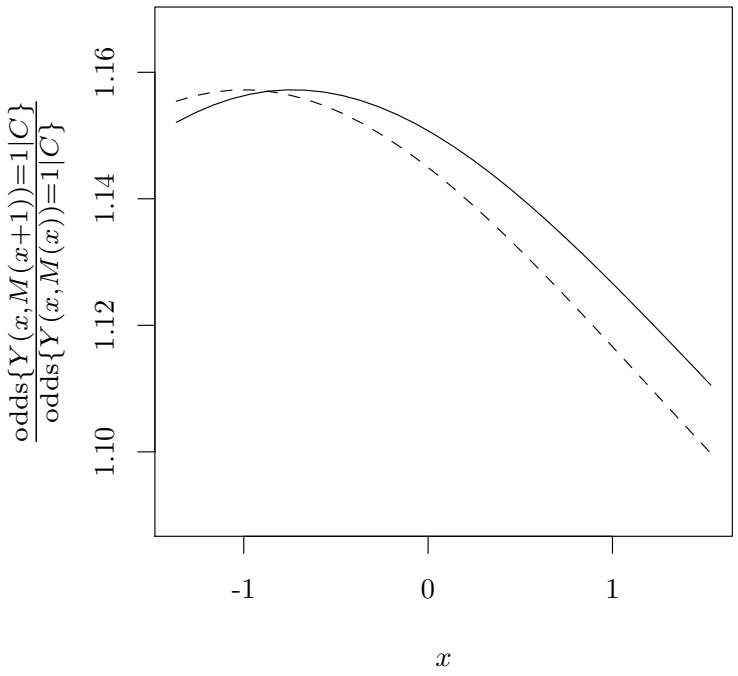

Figure 1: Estimated (total) natural indirect effect odds ratios corresponding to a one-unit change in anxious attachment level as a function of different reference levels for anxious attachment level x (as obtained through direct application of the mediation formula). These are conditional estimates for 43-year-old men (solid curve) and women (dashed curve) with intermediate education levels.

A drawback of direct application of the mediation formula, however, is that combinations of simple models for the mediator and for the outcome often result in complex expressions for natural direct and indirect effects (Lange *et al.* 2012; Vansteelandt *et al.* 2012). For instance, when using logistic regression models

$$\begin{aligned}\text{logit}\,\Pr(M=1|X,C) &= \alpha_0+\alpha_1 X+\alpha_2 C\\ \text{logit}\,\Pr(Y=1|X,M,C) &= \beta_0+\beta_1 X+\beta_2 M+\beta_3 C\end{aligned}$$

for binary mediators and outcomes, the mediation formula yields

$$\begin{aligned}\Pr(Y(x,M(x^*))=1|C) = &\,\text{expit}\,(\beta_0+\beta_1 x+\beta_2+\beta_3 C)\,\text{expit}\,(\alpha_0+\alpha_1 x^*+\alpha_2 C)\\ &+\text{expit}\,(\beta_0+\beta_1 x+\beta_3 C)\,\{1-\text{expit}\,(\alpha_0+\alpha_1 x^*+\alpha_2 C)\}\,,\end{aligned}$$

an expression which depends on exposure and covariate levels in a complicated way. Even though none of the postulated models include interaction terms reflecting effect modification, derived direct and indirect effects estimates will vary with different exposure or covariate levels. This is also illustrated in figure 1, which depicts estimates for the natural indirect effect odds ratio, as obtained by applying the mediation formula to these models fitted to our example dataset (using a dichotomized version of the mediator and baseline covariates C including gender, age and education level). As pointed out before by Lange *et al.* (2012)

and Vansteelandt *et al.* (2012), these convoluted expressions render results difficult to report and hypotheses testing (e.g., testing for moderated mediation) infeasible. As a result, tests of the null hypothesis that the direct or indirect effect is independent of covariates will likely suffer from inflated type I error rates. In certain cases, this complexity can pose a major impediment to routine application of the mediation formula.

Moreover, the **mediation** package only provides natural effect estimates on the additive scale. This may complicate estimation and inference in nonlinear outcome models, mainly when dealing with continuous exposures or covariates, because of induced nonadditivity. Specifically, because the indirect effect is not encoded by a single parameter, but may take on a different value for each level of x, the null hypothesis of no indirect effect over the entire range of exposure levels becomes difficult to test. Similarly, although the **mediation** package enables users to test for effect modification in nonlinear models (i.e., either treatment-mediator interactions or moderated mediation), these hypothesis tests probe research questions in terms of risk differences that are tied to pre-specified exposure or covariate levels. A concern is that these levels might, at least in some applications, need to be chosen in a rather arbitrary way (Loeys *et al.* 2013).

An approach that circumvents the aforementioned complexity but is closely related to application of the mediation formula was proposed recently by Lange *et al.* (2012) and Vansteelandt *et al.* (2012). These authors proposed to directly model the natural effects and introduced a novel class of mean models for nested counterfactuals, which they termed *natural effect models* (also see van der Laan and Petersen 2008, for a similar approach). This approach is implemented in the **medflex** package and provides a viable alternative to the aforementioned software applications because

- it can handle a larger class of parametric models for the mediator and outcome than the software applications that rely on closed-form expressions (refer to Section 4),
- effect estimates can be expressed on more natural scales than the additive scale (i.e., a scale that corresponds to the link-function of the outcome model), thereby avoiding potential (artifactual) dependency on exposure or covariate levels,
- natural effect models simplify testing since the hypotheses of interest can always be captured by one or more model parameters,
- for the most common types of parametric models robust standard errors (based on the sandwich estimator) are available as an alternative to more computer-intensive bootstrap standard errors.

In the next section, we describe this novel class of causal models together with two different approaches that have been suggested in Lange *et al.* (2012) and Vansteelandt *et al.* (2012).

3. Mediation analysis via natural effect models

Natural effect models are conditional mean models for nested counterfactuals $Y(x, M(x^*))$:

$$\mathsf{E}\{Y(x, M(x^*))|C\} = g^{-1}\{\beta' W(x, x^*, C)\}$$

with $g(\cdot)$ a known link function (e.g., the identity or logit link), $W(x, x^*, C)$ a known vector with components that may depend on x, x^* and C, and β a vector including parameters that encode the natural effects of interest.[2] It can, for instance, easily be inferred that in model

$$\mathsf{E}\{Y(x, M(x^*))|C\} = \beta_0 + \beta_1 x + \beta_2 x^* + \beta_3 C,$$

β_1 captures the natural direct effect whereas β_2 captures the natural indirect effect, both corresponding to a one-unit increase in the exposure level. With $g(\cdot)$ the log-link function, for example, the Poisson regression model

$$\log \mathsf{E}\{Y(x, M(x^*))|C\} = \beta_0 + \beta_1 x + \beta_2 x^* + \beta_3 C,$$

enables to quantify the natural direct and indirect effect for count outcomes on a relative and more natural scale. Specifically, in this model, $\exp(\beta_1)$ captures the natural direct effect rate ratio

$$\frac{\mathsf{E}\{Y(x+1, M(x))|C\}}{\mathsf{E}\{Y(x, M(x))|C\}}$$

whereas $\exp(\beta_2)$ captures the natural indirect effect rate ratio

$$\frac{\mathsf{E}\{Y(x, M(x+1))|C\}}{\mathsf{E}\{Y(x, M(x))|C\}},$$

corresponding to a one-unit increase in exposure level. Since each of the effects or quantities of interest are encoded by parameters indexing the natural effect model, the aforementioned limitations related to direct application of the mediation formula can be overcome. As will be illustrated, in nonlinear settings, this facilitates interpretation and hypothesis testing.

3.1. Fitting natural effect models

Before describing the two main approaches for fitting natural effect methods, we first return to our motivating example. The corresponding dataset will then be used to both illustrate these approaches and to demonstrate how they can be implemented in R.

After loading the **medflex** package, displaying the first few rows of the example dataset `UPBdata` provides some insight into the data:

```
R> library(medflex)
R> data(UPBdata)
R> head(UPBdata)
```

```
     att attbin attcat negaff  initiator gender educ age UPB
1  1.001      1      M  0.840     myself      F    M  41   1
2 -0.709      0      L -1.257       both      M    M  42   0
3 -0.709      0      L -1.202       both      F    H  43   0
4  0.606      1      M -0.374 ex-partner      M    H  52   0
5  0.212      1      M  1.945 ex-partner      M    M  32   1
6  2.052      1      H -0.816 ex-partner      M    H  47   0
```

[2] Although Lange *et al.* (2012) primarily described these models as marginal or population-average mean models, throughout this paper we will describe natural effect models as conditional or stratum-specific mean models (i.e., conditional on baseline covariates, as in Vansteelandt *et al.* 2012). A weighting method for fitting marginal or population-average natural effect models is presented in section 7.

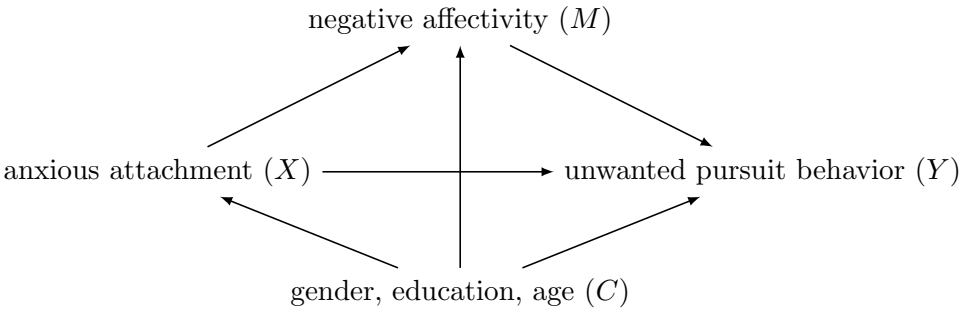

Figure 2: Causal diagram reflecting the mediation hypothesis.

De Smet *et al.* (2012) and Loeys *et al.* (2013) proposed emotional distress or the amount of negative affectivity experienced during the breakup as a mediating variable for the effect of attachment style towards the ex-partner before the breakup on displaying unwanted pursuit behaviors after the breakup. Figure 2 depicts the causal diagram that reflects this mediation hypothesis along with its aforementioned identification assumptions.

As direct and indirect effects are most easily understood for binary exposures, we will use a dichotomized version of anxious attachment level (`attbin`) for didactive purposes. Moreover, negative affectivity (`negaff`) has been standardized to allow for easily interpretable effect estimates. The outcome variable unwanted pursuit behavior (`UPB`) indicates whether (=1) or not (=0) the respondent has engaged in any unwanted pursuit behaviors.

A relatively simple natural effect model is the logistic model

$$\text{logit}\Pr\{Y(x, M(x^*)) = 1|C\} = \beta_0 + \beta_1 x + \beta_2 x^* + \beta_3 C, \tag{1}$$

with x and x^* corresponding to hypothetical levels of the dichotomized version of the anxious attachment variable (i.e., 0 for lower than average or 1 otherwise), $M(x^*)$ corresponding to the level of negative affectivity that would have been reported if anxious attachment level were set to x^*, C a set of baseline covariates, considered sufficient to control for confounding: age (in *years*), gender and education level (`educ`; with `H` or 'high' indicating having obtained at least a bachelor's degree, `M` or 'intermediate' indicating having finished secondary school and `L` or 'low' otherwise), and $Y(x, M(x^*))$ corresponding to the UPB perpetration status that would have been observed if anxious attachment level were set to x and negative affectivity were set to the level that would have been reported if anxious attachment style were set to x^*.

i	X_i	x	x^*	$Y_i(x, M_i(x^*))$
1	1	1	1	Y_1
2	0	0	0	Y_2
⋮	⋮	⋮	⋮	⋮

Table 1: Schematic display of the original dataset.

As an illustration, we schematically display the first two observations in Table 1. For each individual or observation unit i, only the counterfactual outcome $Y_i(X_i, M_i(X_i))$, corresponding

to $Y_i(x, M_i(x^*))$ with x and x^* equal to the observed exposure level X_i, is observed. Postulating a model for nested counterfactuals that encodes both natural direct and indirect effects requires data in which either x or x^* can be kept fixed within each individual while allowing the other variable to vary. Such a procedure amounts to expanding the data along unobserved (x, x^*) combinations. Although, for the data at hand, three (x, x^*) combinations are unobserved for each individual, it is sufficient to introduce only one additional observation corresponding to an unobserved combination for which x does not equal x^* to disentangle natural direct and indirect effects. This data expansion is illustrated in Table 2.

i	X_i	x	x^*	$Y_i(x, M_i(x^*))$
1	1	1	1	Y_1
1	1	1	**0**	.
2	0	0	0	Y_2
2	0	0	**1**	.
⋮	⋮	⋮	⋮	⋮

Table 2: Schematic display of the expanded dataset with missing counterfactual outcomes.

Fitting natural effect models then entails using well-established methods to deal with missingness in the outcome, which results from expanding the data. Throughout, we will describe a weighting- and an imputation-based approach, which, as outlined below, differ mainly in terms of the statistical working models on which they rely (Vansteelandt 2012).

Data expansion is identical for both approaches, but subsequent algorithms for data preparation differ depending on the type of working model. In the **medflex** package, these two steps are implemented in the functions `neWeight` and `neImpute`. Both return an expanded dataset to which the natural effect model can be fitted using the central function `neModel` (see Figure 3). In the next two subsections, we explain both approaches and give example code in R.

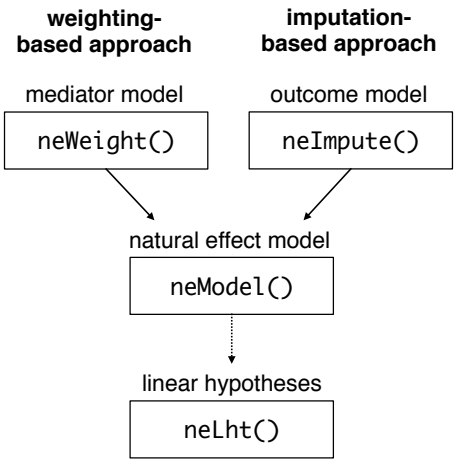

Figure 3: Workflow of the **medflex** package.

3.2. Weighting-based approach

One way to account for missingness in the expanded data is to standardize observed outcomes to the mediator distribution at exposure level x^* rather than the observed level X (which x is set equal to). Building on Hong (2010)'s ratio-of-mediator-probability weighting method, Lange *et al.* (2012) proposed to weight each observation in the expanded dataset by $w_i = p_i(x^*)/p_i(x) = \Pr(M = M_i|X = x^*, C_i)/\Pr(M = M_i|X = x, C_i)$. Estimates for natural direct and indirect effects can then be obtained by regressing the observed outcome on x, x^* and baseline covariates C, weighting each observation in the expanded dataset by its corresponding ratio-of-mediator-probability weight. Intuitively, higher (lower) weights identify observations whose mediator level is more (less) typical for different exposure levels than for the actually observed level, thereby standardizing the observed outcomes to the mediator distribution at exposure levels x^*.[3] This procedure is illustrated in Table 3.

i	X_i	x	x^*	$Y_i(x, M_i(x^*))$	w_i
1	1	1	1	Y_1	1
1	1	1	**0**	Y_1	$\hat{p}_1(\mathbf{0})/\hat{p}_1(1)$
2	0	0	0	Y_2	1
2	0	0	**1**	Y_2	$\hat{p}_2(\mathbf{1})/\hat{p}_2(0)$
⋮	⋮	⋮	⋮	⋮	⋮

Table 3: Schematic display of the weighting-based approach.

Valid estimation of natural effects using the weighting-based approach hinges on adequate specification of the mediator distribution. This can be demanding when exposure or covariates are strongly predictive of the mediator, or when the mediator is continuous, for then even small misspecifications can have a big impact on the weights. In addition to potential weight instability, this forms the main limitation of this approach.

Expanding the data and computing weights for the natural effect model

Using the **medflex** package, expanding the dataset and calculating weights can be done in a single run, using the `neWeight` function. To calculate the weights, a model for the mediator needs to be fitted. For instance, in R, the simple linear model

$$\mathsf{E}(M|X, C) = \alpha_0 + \alpha_1 X + \alpha_2 C,$$

can be fitted using the `glm` function:

```
R> medFit <- glm(negaff ~ factor(attbin) + gender + educ + age,
+                family = gaussian, data = UPBdata)
```

Next, this fitted object needs to be specified as the first argument in `neWeight`, which in turn codes the first predictor variable in the `formula` argument as the exposure and then expands

[3] The interested reader is referred to Appendix A.1, where we give a more technical account on the link between the weighting-based approach and the mediation formula by illustrating that the mediation formula can be rewritten as a weighted mean outcome conditional on x, x^* and C.

the data along hypothetical values of this variable. It is important to note here that, for successful data expansion, categorical exposures should be explicitly coded as factors in the `formula` if they are not yet coded as such in the dataset.

```
R> expData <- neWeight(medFit)
```

Inspecting the first rows of the resulting expanded dataset shows that for each individual two replications have been created:

```
R> head(expData, 4)
```

```
  id attbin0 attbin1    att attcat negaff initiator gender educ age UPB
1  1       1       1  1.001      M   0.84    myself      F    M  41   1
2  1       1       0  1.001      M   0.84    myself      F    M  41   1
3  2       0       0 -0.709      L  -1.26      both      M    M  42   0
4  2       0       1 -0.709      L  -1.26      both      M    M  42   0
```

The new variables `attbin0` and `attbin1` correspond to hypothetical exposure values x and x^*, respectively. By convention, the index '0' is used for parameters (and corresponding auxiliary variables) indexing natural direct effects, whereas the index '1' is used for parameters indexing natural indirect effects in the natural effect model.

To shorten code, one can instead choose to directly specify the `formula`, `family` and `data` arguments in `neWeight`. As illustrated below, this yields identical results:

```
R> expData <- neWeight(negaff ~ factor(attbin) + gender + educ + age,
+                      data = UPBdata)
R> head(expData, 4)
```

```
  id attbin0 attbin1    att attcat negaff initiator gender educ age UPB
1  1       1       1  1.001      M   0.84    myself      F    M  41   1
2  1       1       0  1.001      M   0.84    myself      F    M  41   1
3  2       0       0 -0.709      L  -1.26      both      M    M  42   0
4  2       0       1 -0.709      L  -1.26      both      M    M  42   0
```

By default, `glm` is used as internal model-fitting function. However, other model-fitting functions can be specified in the `FUN` argument (e.g., `vglm` from the **VGAM** package (Yee and Wild 1996)).[4]

Finally, the weights are stored as an attribute of the expanded dataset and can easily be retrieved using the generic `weights` function, e.g., for further inspection of their empirical distribution:

```
R> w <- weights(expData)
R> head(w, 10)
```

[4]In the current version of the package also `vglm` and `vgam` from the **VGAM** package and `gam` from the **gam** package (Hastie 2013) are supported. When specifying model-fitting functions other than `glm` in the `FUN` argument, one might need to specify the `family` argument differently. That is, in a way that is consistent with argument specification of that specific model-fitting function.

```
 [1] 1.000 0.640 1.000 0.494 1.000 0.475 1.000 1.211 1.000 0.326
```

Fitting the natural effect model on the expanded data

After expanding the data and calculating regression weights for each of the replicates, the natural effect model can be fitted using the `neModel` function. Argument specification for this function is similar to that of the `glm` function, which is called internally. However, the `formula` argument now must be specified in function of the variables from the expanded dataset. The latter, in turn, needs to be specified via the `expData` argument. `neModel` automatically extracts the regression weights from this expanded dataset and applies them for model fitting.

Default `glm` standard errors tend to be downwardly biased as the uncertainty inherent to prediction of the weights based on the estimated mediator model is not taken into account. For this reason, `neModel` returns bootstrap standard errors. The number of bootstrap defaults to 1000 and can be set in the `nBoot` argument:

```
R> neMod1 <- neModel(UPB ~ attbin0 + attbin1 + gender + educ + age,
+                    family = binomial("logit"), expData = expData)
```

The `summary` table of the resulting natural effect model object provides these bootstrap standard errors along with corresponding Wald-type z- and p-values.

```
R> summary(neMod1)

Natural effect model
with standard errors based on the non-parametric bootstrap
---
Exposure: attbin
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.74939    1.95012   -0.90  0.36968
attbin01     0.89630    0.31983    2.80  0.00507 **
attbin11     0.40170    0.11358    3.54  0.00041 ***
genderM      0.19516    0.32205    0.61  0.54452
educM       -0.38724    1.86410   -0.21  0.83544
educH       -0.34661    1.87366   -0.18  0.85324
age         -0.00611    0.01571   -0.39  0.69755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As an alternative, robust standard errors based on the sandwich estimator (Liang and Zeger 1986) can be requested by setting `se = "robust"`. Calculation of these standard errors is less computer-intensive and is available for natural effect models with working models fitted via the `glm` function.

```
R> neMod1 <- neModel(UPB ~ attbin0 + attbin1 + gender + educ + age,
+                    family = binomial("logit"), expData = expData,
+                    se = "robust")
R> summary(neMod1)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: attbin
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.74939    0.86469   -2.02  0.04306 *
attbin01     0.89630    0.30456    2.94  0.00325 **
attbin11     0.40170    0.11528    3.48  0.00049 ***
genderM      0.19516    0.30643    0.64  0.52420
educM       -0.38724    0.59399   -0.65  0.51445
educH       -0.34661    0.61101   -0.57  0.57053
age         -0.00611    0.01490   -0.41  0.68193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting model parameters

Exponentiating the model parameter estimates provides estimates that can be interpreted as odds ratios. For instance, for a subject with baseline covariate levels C, altering the level of anxious attachment from low (=0) to high (=1), while controlling negative affectivity at levels as naturally observed for respondents with any given level of anxious attachment x, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}^{\text{NDE}}_{1,0|C} = \frac{\text{odds}\{Y(1, M(x)) = 1|C\}}{\text{odds}\{Y(0, M(x)) = 1|C\}} = \exp(\hat{\beta}_1) = \exp(0.8963) = 2.45.$$

Altering levels of negative affectivity as observed in respondents with low anxious attachment scores to levels that would have been observed if anxious attachment scores of these respondents was high, while controlling their anxious attachment score at any given level x, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}^{\text{NIE}}_{1,0|C} = \frac{\text{odds}\{Y(x, M(1)) = 1|C\}}{\text{odds}\{Y(x, M(0)) = 1|C\}} = \exp(\hat{\beta}_2) = \exp(0.4017) = 1.49.$$

Wald-type confidence intervals can be obtained by applying the `confint` function to the natural effect model object. The confidence level defaults to 95%, but can be changed via the `level` argument. By exponentiating the intervals on the logit scale, we can obtain the corresponding 95% confidence intervals (based on the robust standard errors) on the odds ratio scale:

```
R> exp(confint(neMod1)[c("attbin01", "attbin11"), ])
```

```
          95% LCL 95% UCL
attbin01    1.35    4.45
attbin11    1.19    1.87
```

If standard errors are obtained via the bootstrap procedure, bootstrap confidence intervals are returned. The default type is calculated based on a first order normal approximation (`type = "norm"`), but other types of bootstrap confidence intervals (such as basic bootstrap, bootstrap percentile and bias-corrected and accelerated confidence intervals) can be obtained by setting the `type` argument to the desired type.[5]

3.3. Imputation-based approach

The second approach avoids reliance on a model for the mediator distribution and instead requires fitting a working model for the outcome mean (Vansteelandt *et al.* 2012). By setting x^* (rather than x) equal to the observed exposure level X, unobserved nested counterfactuals can be imputed using any appropriate mean model for the outcome. That is, since the counterfactual mediator level $M(x^*)$ equals the observed mediator level M within the stratum of individuals with observed exposure level $X = x^*$, $Y(x, M(x^*))$ equals $Y(x, M)$. The latter can then be imputed using fitted values $\hat{\mathsf{E}}(Y|X = x, M, C)$ based on an appropriate model for the outcome mean, henceforth referred to as the imputation model, with exposure level X set to x and with mediator M and baseline covariates C set to their observed values. Finally, natural direct and indirect effect estimates can be obtained upon fitting a natural effect model to the imputed dataset.[6] This procedure is illustrated in Table 4. For ease of implementation, observed nested counterfactuals are imputed as well in the **medflex** package.[7]

i	X_i	x	x^*	$Y_i(x, M_i(x^*))$
1	1	1	1	Y_1
1	1	**0**	1	$\hat{Y}_1(\mathbf{0}, M_1)$
2	0	0	0	Y_2
2	0	**1**	0	$\hat{Y}_2(\mathbf{1}, M_2)$
⋮	⋮	⋮	⋮	⋮

Table 4: Schematic display of the imputation-based approach. $\hat{Y}_i(x, M_i)$ represent the imputed counterfactual outcomes.

Although circumventing stability issues inherent to weighting, the imputation-based approach does not come without limitations. As in other imputation settings, one must pay due attention to coherent model specification between the imputer's model and the analyst's model (i.e., in this case, the natural effect model). To this end, Vansteelandt *et al.* (2012) and Loeys

[5]The `type` argument in `confint` corresponds to that of the `boot.ci` function from the **boot** package (Canty and Ripley 2014), which is called internally.

[6]In Appendix A.2, we demonstrate the link between the mediation formula and the imputation-based approach to natural effect models by showing that the mediation formula can be rewritten as a formulation that prescribes estimating nested counterfactuals by calculating the mean of imputed nested counterfactuals, conditional on x, x^* and C.

[7]Simulation studies (not shown here) have shown that this procedure does not lead to bias or loss of efficiency.

et al. (2013) advocated the use of a rich imputation model to reduce the impact of model incongeniality in terms of misspecification bias. In this vein, the **medflex** package also allows users to fit the imputation model using machine learning techniques, such as the ensemble learner as implemented in the **SuperLearner** package (Polley and van der Laan 2014).[8]

Expanding the data and imputing nested counterfactuals

Although application of the imputation-based approach is similar to that of the weighting-based approach, it differs in some key respects. These differences are mainly captured by differences between the functions `neWeight` and `neImpute`. Argument specification of this function is identical to that of `neWeight`, unless indicated otherwise.

As for the weighted-based approach, the first step amounts to fitting a working model. Instead of a model for the mediator, the imputation-based approach requires fitting a mean model for the outcome. Moreover, this model should at least reflect the structure of natural effect model (1), to avoid the aforementioned lack of a coherent model specification. That is, it should at least contain all effects of the natural effect model with x^* replaced by M. For instance, a simple logistic regression model

$$\text{logit}\,\Pr(Y = 1|X, M, C) = \gamma_0 + \gamma_1 X + \gamma_2 M + \gamma_3 C,$$

can be fitted in R using the `glm` function:

```
R> impFit <- glm(UPB ~ factor(attbin) + negaff + gender + educ + age,
+                family = binomial("logit"), data = UPBdata)
```

In order for `neImpute` to identify the predictor variables in the `formula` argument correctly as either exposure, mediator(s) or baseline covariates, they need to be entered in a particular order. That is, the first predictor variable again needs to point to the exposure and the second to the mediator, irrespective of the use of operators (i.e., `+`, `*` and `:`). All other predictors are automatically coded as baseline covariates.

This fitted object then needs to be entered as the first argument in `neImpute`:

```
R> expData <- neImpute(impFit)
```

Alternatively, the `formula`, `family` and `data` arguments can be directly specified in `neImpute`:

```
R> expData <- neImpute(UPB ~ factor(attbin) + negaff + gender + educ + age,
+                      family = binomial("logit"), data = UPBdata)
```

Similar to `neWeight`, `neImpute` first expands the data along hypothetical exposure values. Instead of calculating weights for these new observations, `neImpute` then imputes the nested counterfactual outcomes by fitted values based on the imputation model. As illustrated below, the resulting expanded dataset includes two imputed nested counterfactual outcomes for each subject:

```
R> head(expData, 4)
```

[8] An example is given in the help files of the package and can be consulted via `?neImpute.default`. Only bootstrap standard errors are available when fitting the imputation model using the `SuperLearner` function.

```
  id attbin0 attbin1    att attcat negaff initiator gender educ age    UPB
1  1       1       1  1.001      M   0.84    myself      F    M  41 0.3085
2  1       0       1  1.001      M   0.84    myself      F    M  41 0.1442
3  2       0       0 -0.709      L  -1.26      both      M    M  42 0.0522
4  2       1       0 -0.709      L  -1.26      both      M    M  42 0.1272
```

Fitting the natural effect model on the imputed data

After expanding and imputing the data, specifying the natural effect model can be done as for the weighting-based approach:

```
R> neMod1 <- neModel(UPB ~ attbin0 + attbin1 + gender + educ + age,
+                    family = binomial("logit"), expData = expData,
+                    se = "robust")
```

Again, bootstrap or robust standard errors are reported in the output of the `summary` function, in order to account for the uncertainty inherent to the working model (i.e., in this case, the imputation model):

```
R> summary(neMod1)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: attbin
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.80711    0.82933   -2.18  0.02933 *
attbin01     0.90796    0.28959    3.14  0.00172 **
attbin11     0.37639    0.10055    3.74  0.00018 ***
genderM      0.22792    0.28698    0.79  0.42708
educM       -0.21261    0.54347   -0.39  0.69565
educH       -0.27774    0.55336   -0.50  0.61573
age         -0.00728    0.01489   -0.49  0.62493
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Natural direct and indirect effect odds ratio estimates and their confidence intervals can be obtained as before.

4. Dealing with different types of variables

In the previous section, we used a dichotomized version of the continuous exposure variable `att`. However, the natural effect model framework easily extends to different types of exposure, mediator or outcome variables. In the following two subsections, we give a detailed

	outcome type					
	binary		count		continuous	
mediator type	`neWeight`	`neImpute`	`neWeight`	`neImpute`	`neWeight`	`neImpute`
binary	✓	✓	✓	✓	✓	✓
count	✓	✓	✓	✓	✓	✓
continuous	✓	✓	✓	✓	✓	✓
ordinal		✓		✓		✓
nominal	✓*	✓	✓*	✓	✓*	✓

Table 5: Types of variables that can be dealt with in the **medflex** package. Natural effect models are currently restricted to models that can be fitted with the `glm` function. '*' indicates that robust standard errors are not available.

description on how to fit natural effect models with multicategorical (i.e., ordinal or nominal) and continuous exposures. In these subsections, as well as throughout the remainder of this paper, we will focus on the imputation-based approach when introducing new features of the **medflex** package. Unless indicated otherwise, the weighting-based approach can be applied analogously.

An overview of the types of mediators and outcomes the **medflex** package can currently handle, is given in Table 5. When using the weighting-based approach, models for binary, count and continuous mediators can be fitted using the `glm` function or the `vglm` function from the **VGAM** package. Models for nominal mediators, on the other hand, can only be fitted using the `vglm` function (setting `family = multinomial`).[9] Although models for ordinal mediators are not compatible with the `neWeight` function, ordered factors can easily be treated as nominal variables. Finally, the imputation-based approach can deal with virtually any type of mediator as it does not require the specification of a mediator model.

4.1. Multicategorical exposures

Methods for dealing with multicategorical treatments or exposures, as encountered in e.g., multiple intervention studies, in which multiple experimental conditions are compared to a control condition, have rarely been described within the mediation literature (although see Hayes and Preacher 2014; Tingley *et al.* 2014, for some notable exceptions).

In this section, we illustrate how to expand the dataset and fit natural effect models when using a multicategorical exposure. In this example, instead of using the binary exposure variable `attbin`, we use a discretized version of anxious attachment style, named `attcat` (with `L` indicating low, `M` indicating intermediate and `H` indicating high anxious attachment levels).

Inspecting the first rows of the expanded dataset shows that the number of replications for each subject again corresponds to the number of unique levels of the categorical exposure variable. That is, the auxiliary variable x^* (`attcat1`) is fixed to the observed exposure level,

[9] In the current version of the package, when using working models for weighting (either when adopting the weighting-based approach or when fitting population-average natural effect models), robust standard errors are only available if these working models are fitted using `glm` and their outcomes (i.e., either an exposure or a mediator) follow either a normal, binomial or Poisson distribution.

whereas the other, x (`attcat0`), enumerates all potential exposure levels.

```
R> expData <- neImpute(UPB ~ attcat + negaff + gender + educ + age,
+                       family = binomial, data = UPBdata)
R> head(expData)

  id attcat0 attcat1    att attbin negaff initiator gender educ age    UPB
1  1       M       M  1.001      1   0.84    myself      F    M  41 0.2842
2  1       H       M  1.001      1   0.84    myself      F    M  41 0.3593
3  1       L       M  1.001      1   0.84    myself      F    M  41 0.1329
4  2       L       L -0.709      0  -1.26      both      M    M  42 0.0501
5  2       M       L -0.709      0  -1.26      both      M    M  42 0.1202
6  2       H       L -0.709      0  -1.26      both      M    M  42 0.1618
```

The `summary` table returns estimates for the natural direct and indirect effect log odds ratios comparing intermediate and high anxious attachment levels to low levels of anxious attachment (i.e., the reference level). The `neEffdecomp` function, described in section 6.2, can be used to derive (log) odds ratios corresponding to the contrast between high and intermediate levels of anxious attachment.

```
R> neMod <- neModel(UPB ~ attcat0 + attcat1 + gender + educ + age,
+                    family = binomial, expData = expData, se = "robust")
R> summary(neMod)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: attcat
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.86733    0.85171   -2.19  0.02835 *
attcat0M     0.89868    0.32516    2.76  0.00571 **
attcat0H     1.21911    0.37591    3.24  0.00118 **
attcat1M     0.32849    0.09549    3.44  0.00058 ***
attcat1H     0.57097    0.15517    3.68  0.00023 ***
genderM      0.20343    0.28548    0.71  0.47610
educM       -0.19334    0.53231   -0.36  0.71645
educH       -0.30124    0.54530   -0.55  0.58066
age         -0.00894    0.01532   -0.58  0.55979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overall assessment of natural effects (i.e., a joint comparison of all levels of the exposure) cannot be based on the default `summary` output, but instead requires an Anova table for the natural effect model, which can be obtained using the `Anova` function from the **car** package (Fox and Weisberg 2011):

```
R> library(car)
R> Anova(neMod)

Analysis of Deviance Table (Type II tests)

Response: UPB
           Df Chisq Pr(>Chisq)
attcat0     2 11.66    0.00293 **
attcat1     2 15.42    0.00045 ***
gender      1  0.51    0.47610
educ        2  0.35    0.83781
age         1  0.34    0.55979
Residuals 1146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both type-II (the default) and type-III Anova tables can be requested by specifying the desired type via the `type` argument. This table includes corresponding Wald χ^2 tests for multivariate hypotheses which account for the uncertainty inherent to the working model. The output suggests that the natural direct and indirect effect odds differ significantly between the three exposure levels.

4.2. Continuous exposures

In contrast to the **mediation** package, hypothesis testing for natural direct and indirect effects along the entire support of continuous exposures is facilitated by defining causal effects on their most natural scale. In this section, we use the continuous variable `att`, a standardized version of the original anxious attachment variable.

For continuous variables, expanding the dataset along unobserved (x, x^*) combinations requires a slightly adapted approach than for categorical exposures. Instead of enumerating all the levels of the exposure to construct auxiliary variables x and x^* for each subject, Vansteelandt *et al.* (2012) proposed to draw specific quantiles from the conditional distribution of the exposure given baseline covariates. By default, these hypothetical exposure levels are drawn from a linear model for the exposure, conditional on a linear combination of all covariates specified in the working model.[10]

Both `neWeight` and `neImpute` allow to choose the number of draws to sample from this conditional distribution via the `nRep` argument (which defaults to 5).

```
R> expData <- neImpute(UPB ~ att + negaff + gender + educ + age,
+                      family = binomial("logit"), data = UPBdata, nRep = 3)
R> head(expData)

  id      att0   att1 attbin attcat negaff initiator gender educ age    UPB
1  1 -1.64e+00  1.001      1      M   0.84    myself      F    M  41 0.1029
```

[10] If one wishes to use another model for the exposure, this default model specification can be overruled by referring to a fitted model object in the `xFit` argument. Misspecification of this sampling model does not induce bias in the estimated coefficients and standard errors of the natural effect model.

```
2  1  8.02e-06  1.001      1      M   0.84    myself      F    M  41 0.2108
3  1  1.64e+00  1.001      1      M   0.84    myself      F    M  41 0.3834
4  2 -1.66e+00 -0.709      0      L  -1.26      both      M    M  42 0.0373
5  2 -1.82e-02 -0.709      0      L  -1.26      both      M    M  42 0.0827
6  2  1.63e+00 -0.709      0      L  -1.26      both      M    M  42 0.1734
```

Specification of the natural effect model via `neModel` can be done as described before:

```
R> neMod1 <- neModel(UPB ~ att0 + att1 + gender + educ + age,
+                    family = binomial("logit"), expData = expData,
+                    se = "robust")
R> summary(neMod1)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: att
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.99234    0.81485   -1.22  0.22329
att0         0.48375    0.13130    3.68  0.00023 ***
att1         0.22172    0.05714    3.88  0.00010 ***
genderM      0.18479    0.28656    0.64  0.51900
educM       -0.33794    0.54320   -0.62  0.53386
educH       -0.43704    0.55191   -0.79  0.42843
age         -0.00894    0.01539   -0.58  0.56135
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output illustrates that defining natural effects on the (log) odds ratio scale allows to capture each of these effects along the entire support of the exposure by a single parameter. For instance, for a subject with baseline covariate levels C, the direct and indirect effects of one standard deviation increase in anxious attachment level (i.e., from x to $x+1$) correspond to an increase in the odds of displaying unwanted pursuit behaviors by a factor

$$\widehat{\text{OR}}^{\text{NDE}}_{x+1,x|C} = \frac{\text{odds}\left\{Y(x+1, M(x)) = 1|C\right\}}{\text{odds}\left\{Y(x, M(x)) = 1|C\right\}} = \exp(\hat{\beta}_1) = \exp(0.48) = 1.62,$$

and

$$\widehat{\text{OR}}^{\text{NIE}}_{x+1,x|C} = \frac{\text{odds}\left\{Y(x, M(x+1)) = 1|C\right\}}{\text{odds}\left\{Y(x, M(x)) = 1|C\right\}} = \exp(\hat{\beta}_2) = \exp(0.22) = 1.25,$$

respectively, regardless of the initial level x. Defining natural effects on the risk difference scale (as in the **medflex** package) would not have enabled to capture these by a single parameter along the entire support of the exposure, because of induced non-additivity (an artificial example illustrating this induced non-additivity is given in Figure 4 of Loeys *et al.* 2013).

Throughout the remainder of the paper, we will continue to use the original continuous exposure variable, `att`.

5. Effect modification of natural effects

5.1. Exposure-mediator interactions: relaxing the no interaction assumption

So far, the considered natural effect models reflected the assumption that the exposure and mediator do not interact in their effect on the outcome. In particular, the natural direct effect odds ratio

$$\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(x) = \frac{\mathrm{odds}\left\{Y(1, M(x)) = 1|C\right\}}{\mathrm{odds}\left\{Y(0, M(x)) = 1|C\right\}}$$

was declared to be the same for each choice of mediator level $M(x)$, and hence for each choice of x, at which the mediator is evaluated, while, similarly, the natural indirect effect odds ratio

$$\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(x) = \frac{\mathrm{odds}\left\{Y(x, M(1)) = 1|C\right\}}{\mathrm{odds}\left\{Y(x, M(0)) = 1|C\right\}}$$

was declared to be the same for each choice of exposure level x at which the outcome was evaluated. In other words, the effects Robins and Greenland (1992) referred to as the *pure* direct effect, $\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0)$, and *total* direct effect, $\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(1)$, were assumed to be equal. Likewise, the *pure* indirect effect, $\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0)$, and *total* indirect effect, $\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(1)$, were assumed to be equal. In many studies, these assumptions may not a priori be plausible.

As pointed out by VanderWeele (2013), total causal effects can be decomposed into a pure direct effect, a pure indirect effect and a mediated interactive effect. On an additive scale, the latter can be described as either the difference between total direct and pure direct effects or as the difference between total indirect and pure indirect effects. Similarly, within the natural effect model framework, the total effect odds ratio

$$\mathrm{OR}_{1,0|C} \equiv \frac{\mathrm{odds}\left\{Y(1, M(1)) = 1|C\right\}}{\mathrm{odds}\left\{Y(0, M(0)) = 1|C\right\}}$$

can be expressed as the product of the pure direct and pure indirect effect odds ratios and the mediated interaction odds ratio

$$\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0) \cdot \mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0) \cdot \frac{\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(1)}{\mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0)} = \mathrm{OR}^{\mathrm{NDE}}_{1,0|C}(0) \cdot \mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0) \cdot \frac{\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(1)}{\mathrm{OR}^{\mathrm{NIE}}_{1,0|C}(0)}.$$

Rather than reflecting the *difference* between total and pure direct or indirect effects, the mediated interaction odds ratio corresponds to the *ratio* of total and pure direct or indirect effect odds ratios.

In a logistic natural effect model, testing for exposure-mediator interaction amounts to testing whether the mediated interaction odds ratio differs from 1, or equivalently, on the scale of the linear predictor, whether the corresponding log odds ratio, as captured by β'_3 in natural effect model

$$\mathrm{logit}\,\Pr\left\{Y(x, M(x^*)) = 1|C\right\} = \beta'_0 + \beta'_1 x + \beta'_2 x^* + \beta'_3 x \cdot x^* + \beta'_4 C, \tag{2}$$

differs from 0. When including this interaction term in the outcome model, β_1' and β_2' will index the pure direct and indirect effect log odds ratios, respectively.

When applying the imputation-based approach, the working model needs to at least reflect the structure of the final natural effect model (as has been pointed out in Section 3.3). This requires the user to first (re)fit the imputation model accordingly. For instance, the minimal imputation model for natural effect model (2) would be the logistic regression model

$$\text{logit}\Pr(Y = 1|X, M, C) = \gamma_0' + \gamma_1' X + \gamma_2' M + \gamma_3' X \cdot M + \gamma_4' C.$$

The output of the corresponding natural effect model object suggests there is no evidence for mediated interaction at the 5% significance level.

```
R> expData <- neImpute(UPB ~ att * negaff + gender + educ + age,
+                      family = binomial("logit"), data = UPBdata)
R> neMod2 <- neModel(UPB ~ att0 * att1 + gender + educ + age,
+                    family = binomial("logit"), expData = expData,
+                    se = "robust")
R> summary(neMod2)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: att
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9040     0.8036   -1.12   0.2606
att0          0.4790     0.1326    3.61   0.0003 ***
att1          0.1824     0.0599    3.05   0.0023 **
genderM       0.1889     0.2873    0.66   0.5108
educM        -0.4200     0.5409   -0.78   0.4375
educH        -0.5067     0.5497   -0.92   0.3566
age          -0.0110     0.0152   -0.72   0.4693
att0:att1     0.0788     0.0504    1.56   0.1177
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.2. Effect modification by baseline covariates

One might additionally wish to determine whether direct or indirect effects generalize across different strata of the population and across different conditions.

In our example, researchers might for instance investigate whether the extent to which the effect of anxious attachment level on engaging in UPBs is mediated through the experience of negative affectivity differs between men and women or between people with different education levels (Muller, Judd, and Yzerbyt 2005; Preacher, Rucker, and Hayes 2007). In the scope

of natural effect models, this moderation mediation hypothesis can be probed by allowing the conditional indirect effect, as indexed by β_2 in model (1), to depend on gender, C_1, as expressed in model (3) below

$$\text{logit}\Pr\{Y(x, M(x^*)) = 1|C\} = \beta_0'' + \beta_1''x + \beta_2''x^* + \beta_3''x^* \cdot C_1 + \beta_4''C, \tag{3}$$

in which testing whether $\beta_3'' = 0$ corresponds to testing for moderated mediation by gender.

```
R> impData <- neImpute(UPB ~ (att + negaff) * gender + educ + age,
+                       family = binomial("logit"), data = UPBdata)
R> neMod3 <- neModel(UPB ~ att0 + att1 * gender + educ + age,
+                     family = binomial("logit"), expData = impData,
+                     se = "robust")
R> summary(neMod3)


Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: att
Mediator(s): negaff
---
Parameter estimates:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.00323    0.81391   -1.23  0.21772
att0          0.48405    0.13080    3.70  0.00021 ***
att1          0.21077    0.07683    2.74  0.00608 **
genderM       0.17245    0.28897    0.60  0.55066
educM        -0.34572    0.54383   -0.64  0.52496
educH        -0.44891    0.55460   -0.81  0.41827
age          -0.00833    0.01547   -0.54  0.59050
att1:genderM  0.03583    0.12195    0.29  0.76894
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output suggests that the natural indirect effect does not differ significantly between men and women.

In a similar way, researchers can probe effect modification by education level. Suppose, for instance, that one wishes to test whether education level moderates both the direct and indirect effect. This can be done by fitting the natural effect model

$$\begin{aligned}\text{logit}\Pr\{Y(x, M(x^*)) = 1|C\} = \beta_0^\star + \beta_1^\star x + \beta_2^\star x^* + \beta_3^\star x \cdot C_{2,1} + \beta_4^\star x \cdot C_{2,2} \\ + \beta_5^\star x^* \cdot C_{2,1} + \beta_6^\star x^* \cdot C_{2,2} + \beta_7^\star C,\end{aligned} \tag{4}$$

with $C_{2,1}$ and $C_{2,2}$ dummy variables encoding the three education levels. Effect modification of the natural indirect (direct) effect by education level in model (4) is then captured by $\beta_5^\star$ and $\beta_6^\star$ ($\beta_3^\star$ and $\beta_4^\star$).

```
R> impData <- neImpute(UPB ~ (att + negaff) * educ + gender + age,
+                       family = binomial("logit"), data = UPBdata)
R> neMod4 <- neModel(UPB ~ (att0 + att1) * educ + gender + age,
+                     family = binomial("logit"), expData = impData,
+                     se = "robust")
```

Testing for moderation by a multicategorical variable calls for a multivariate test, which can again be obtained by requesting an Anova table for the natural effect model.

6. Tools for deriving and visualizing causal effect estimates

In this section, we highlight tools that can aid in deriving and visualizing specific causal effect estimates of interest. These tools might prove to be useful for gaining insight, especially for more complex models including interaction terms involving natural effect parameters.

6.1. Linear combinations of parameter estimates

Although effect estimates for e.g., the total causal effect can easily be derived from the `summary` table of a natural effect model, its standard error and confidence interval cannot. To this end, the function `neLht`, which exploits the functionality of the `glht` function from the **multcomp** package (Hothorn, Bretz, and Westfall 2008) can be of use. This function enables the calculation of linear combinations of parameter estimates as well as their corresponding standard errors and confidence intervals based on the bootstrap or robust variance-covariance matrix of the natural effect model.

For instance, in model (2), the total direct and indirect effect can be expressed on the log odds scale as $\hat{\beta}'_1 + \hat{\beta}'_3$ and $\hat{\beta}'_2 + \hat{\beta}'_3$, respectively. Similarly, the total causal effect log odds ratio is captured by $\hat{\beta}'_1 + \hat{\beta}'_2 + \hat{\beta}'_3$. As the argument for the linear function, `linfct`, needs to be specified in terms of one or more linear hypotheses, these effects can be specified as illustrated below:

```
R> lht <- neLht(neMod2, linfct = c("att0 + att0:att1 = 0",
+                                  "att1 + att0:att1 = 0",
+                                  "att0 + att1 + att0:att1 = 0"))
```

The corresponding odds ratios and their confidence intervals can be requested by exponentiating the coefficients and confidence intervals of the resulting object:

```
R> exp(cbind(coef(lht), confint(lht)))
```

```
                            95% LCL 95% UCL
att0 + att0:att1        1.75    1.33    2.29
att1 + att0:att1        1.30    1.14    1.47
att0 + att1 + att0:att1 2.10    1.62    2.72
```

Separate univariate tests for linear hypothesis objects can be requested using the `summary` function:

```
R> summary(lht)

Linear hypotheses for natural effect models
with standard errors based on the sandwich estimator
---
                         Estimate Std. Error z value Pr(>|z|)
att0 + att0:att1           0.5578     0.1376    4.05  5.1e-05 ***
att1 + att0:att1           0.2613     0.0645    4.05  5.2e-05 ***
att0 + att1 + att0:att1    0.7403     0.1330    5.56  2.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Univariate p-values reported)
```

In contrast to the `summary` table for `glht` objects, which yields p-values that are adjusted for multiple testing, tests returned by the `summary` function applied to `neLht` objects report unadjusted univariate tests. Adjusted tests can be obtained by setting `test = adjusted()` (for more details consult the help page of the `adjusted()` function from the **multcomp** package (Hothorn *et al.* 2008)).

6.2. Effect decomposition

If one is only interested in the natural effect parameters, the convenience function `neEffdecomp` can be used instead of `neLht`. This function automatically retains the natural effect estimates and generates a linear hypothesis object that reflects the most suitable effect decomposition:

```
R> effdecomp <- neEffdecomp(neMod2)
R> summary(effdecomp)

Effect decomposition on the scale of the linear predictor
with standard errors based on the sandwich estimator
---
conditional on: gender, educ, age
with x* = 0, x = 1
---
                      Estimate Std. Error z value Pr(>|z|)
pure direct effect      0.4790     0.1326    3.61   0.0003 ***
total direct effect     0.5578     0.1376    4.05  5.1e-05 ***
pure indirect effect    0.1824     0.0599    3.05   0.0023 **
total indirect effect   0.2613     0.0645    4.05  5.2e-05 ***
total effect            0.7403     0.1330    5.56  2.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Univariate p-values reported)
```

By default, reference levels for the exposure, x and x^*, are chosen to be 1 and 0, respectively. If one wishes to evaluate causal effects at different reference levels (e.g., if the natural effect model allows for mediated interaction or if it includes quadratic or higher-order polynomial

terms for the exposure), these can be specified as a vector of the form `c(x*,x)` via the `xRef` argument.

The output indicates that, for a subject with baseline covariate levels C, a standard deviation increase from the average level of anxious attachment (=0), increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}^{\text{NDE}}_{1,0|C}(0) = \frac{\text{odds}\{Y(1,M(0))=1|C\}}{\text{odds}\{Y(0,M(0))=1|C\}} = \exp(\hat{\beta}'_1) = 1.61$$

when controlling negative affectivity at levels as naturally observed for respondents with average anxious attachment levels, or with a factor

$$\widehat{\text{OR}}^{\text{NDE}}_{1,0|C}(1) = \frac{\text{odds}\{Y(1,M(1))=1|C\}}{\text{odds}\{Y(0,M(1))=1|C\}} = \exp(\hat{\beta}'_1 + \hat{\beta}'_3) = 1.75$$

when controlling negative affectivity at levels as naturally observed for respondents with anxious attachment levels one standard deviation above the average level.

On the other hand, altering levels of negative affectivity as observed in respondents with average levels of anxious attachment to levels that would have been observed if anxious attachment scores of these respondents increased with a standard deviation, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}^{\text{NIE}}_{1,0|C}(0) = \frac{\text{odds}\{Y(0,M(1))=1|C\}}{\text{odds}\{Y(0,M(0))=1|C\}} = \exp(\hat{\beta}'_2) = 1.20$$

when controlling their anxious attachment level at the average, or with a factor

$$\widehat{\text{OR}}^{\text{NIE}}_{1,0|C}(1) = \frac{\text{odds}\{Y(1,M(1))=1|C\}}{\text{odds}\{Y(1,M(0))=1|C\}} = \exp(\hat{\beta}'_2 + \hat{\beta}'_3) = 1.30$$

when controlling their anxious attachment level one standard deviation above the average.

The total causal effect odds ratio can be expressed as the product of the pure direct and indirect effect odds ratios and the mediated interaction odds ratio: a standard deviation increase from the average level of anxious attachment approximately doubles the odds of displaying unwanted pursuit behaviors.

$$\widehat{\text{OR}}_{1,0|C} = \frac{\text{odds}\{Y(1,M(1))=1|C\}}{\text{odds}\{Y(0,M(0))=1|C\}} = \exp(\hat{\beta}'_1 + \hat{\beta}'_2 + \hat{\beta}'_3) = 2.10.$$

If the model includes terms reflecting effect modification by baseline covariates (e.g., as in model (3)), effect decomposition is by default evaluated at covariate levels that correspond to 0 for continuous covariates and to the reference level for categorical covariates coded as factors. However, for this type of models, it might often be insightful to evaluate natural effect components at different covariate levels than the default levels. This can be done via the `covLev` argument, which requires a vector including valid levels for modifier covariates specified in the natural effect model. An example of effect decomposition for women (`gender = "F"`, the default covariate level) and men (`gender = "M"`) in model (3) is given in the R code below:

```
R> neEffdecomp(neMod3)
```

```
Effect decomposition on the scale of the linear predictor
---
conditional on: gender = F, educ, age
with x* = 0, x = 1
---
                        Estimate
natural direct effect      0.484
natural indirect effect    0.211
total effect               0.695
```

```
R> neEffdecomp(neMod3, covLev = c(gender = "M"))
```

```
Effect decomposition on the scale of the linear predictor
---
conditional on: gender = M, educ, age
with x* = 0, x = 1
---
                        Estimate
natural direct effect      0.484
natural indirect effect    0.247
total effect               0.731
```

6.3. Global hypothesis tests

Global hypothesis tests considering all linear hypothesis simultaneously can be requested by specifying `test = Chisqtest()`. For instance, in model (4), instead of using the `Anova` function, one could also test for moderated mediation by means of a global hypothesis test involving the relevant parameters $\beta_5^\star$ and $\beta_6^\star$:

```
R> modmed <- neLht(neMod4, linfct = c("att1:educM = 0", "att1:educH = 0"))
R> summary(modmed, test = Chisqtest())
```

```
Global linear hypothesis test for natural effect models
with standard errors based on the sandwich estimator
---
  Chisq DF Pr(>Chisq)
1   3.8  2      0.149
```

6.4. Visualizing effect estimates and their uncertainty

Finally, the generic `plot` function can be applied to linear hypothesis objects to visualize (linear combinations of) effect estimates and their uncertainty by means of confidence interval plots. To obtain estimates and confidence intervals on the odds ratio scale, one can specify `transf = exp` in order to exponentiate the original parameter estimates (on the log odds ratio scale).

Applying the `plot` function to a natural effect model object automatically retains the causal effect estimates of interest, generates a linear hypothesis object using `neEffdecomp` and then plots its corresponding estimates and confidence intervals, as shown in Figure 4.

```
R> par(mfrow = c(1, 2))
R> plot(neMod2, xlab = "log odds ratio")
R> plot(neMod2, xlab = "odds ratio", transf = exp)
```

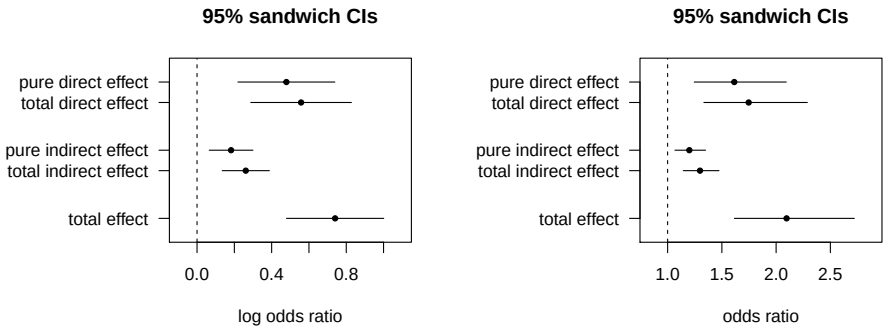

Figure 4: Effect decomposition on the log odds ratio and odds ratio scales.

The default exposure reference and covariate levels for these plots are the same as for the `neEffdecomp` function, but can again be altered via the corresponding arguments `xRef` and `covLev`.

7. Population-average natural effects

In all previous sections, we defined natural effects as conditional or stratum-specific effects (i.e., conditional on baseline covariates). However, the **medflex** package also allows to estimate population-average natural effects. As demonstrated in Appendix A.3 and A.4, rewriting the mediation formula reveals that estimation of these population-average effects requires weighting by the reciprocal of the conditional exposure distribution in order to adjust for confounding (also see Albert 2012; Vansteelandt 2012).

As a consequence, a model for the exposure distribution needs to be fitted and specified as an additional working model, e.g.,

```
R> expFit <- glm(att ~ gender + educ + age, data = UPBdata)
```

Since specifying population-average natural effect models using the `neModel` is equivalent for the weighting- and imputation-based approaches, in the remainder of this section, we demonstrate how to proceed when adhering to the imputation-based approach. Moreover, when estimating population-average natural effects, incoherence between imputation and natural effect models is less of a concern as the latter does not require modeling the relation between outcome and covariates. The (first) working model can again be fitted using the same commands as before:

```
R> impData <- neImpute(UPB ~ att + negaff + gender + educ + age,
+                      family = binomial("logit"), data = UPBdata)
```

Each observation in the expanded dataset to which the marginal natural effect model indexing the population-average natural direct and indirect effects

$$\text{logit}\Pr\{Y(x, M(x^*)) = 1\} = \theta_0 + \theta_1 x + \theta_2 x^* \tag{5}$$

is fitted, needs to be weighted by the reciprocal of the exposure probability density, $\Pr(X|C)$, evaluated at the observed exposure. The fitted model object that is used to calculate regression weights needs to be specified in the `xFit` argument of the `neModel` function:

```
R> neMod5 <- neModel(UPB ~ att0 + att1, family = binomial("logit"),
+                    expData = impData, xFit = expFit, se = "robust")
R> summary(neMod5)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: att
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6451     0.1471  -11.18  < 2e-16 ***
att0          0.4756     0.1298    3.66  0.00025 ***
att1          0.2439     0.0711    3.43  0.00061 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both the marginal natural direct and indirect effect odds ratios again seem to be significantly different from 1: increasing the anxious attachment level from average to one standard error above average, while keeping negative affectivity fixed at levels that would naturally have been reported had anxious attachment level been fixed at any given level x^*, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\text{OR}}_{1,0}^{\text{NDE}} = \frac{\text{odds}\{Y(1, M(x^*)) = 1\}}{\text{odds}\{Y(0, M(x^*)) = 1\}} = \exp(\hat{\theta}_1) = 1.61.$$

A similar interpretation can again be made for the natural indirect effect.

8. Multiple mediators: a joint mediator approach

In many settings multiple mediators may be of interest. In our example, one could argue that being anxiously attached to one's partner makes respondents more hesitant to end their relationship and that, in turn, not having initiated the break-up causes them to engage in unwanted pursuit behaviors more often. In this sense, initiator status (`initiator`: either `"both"`, `"ex-partner"`, or `"myself"`) can also be considered as a mediator, which we denote L.

If hypothesized mediators do not affect one another, one can fit separate natural effect models (each with a different working model involving only one of the mediators) to assess the

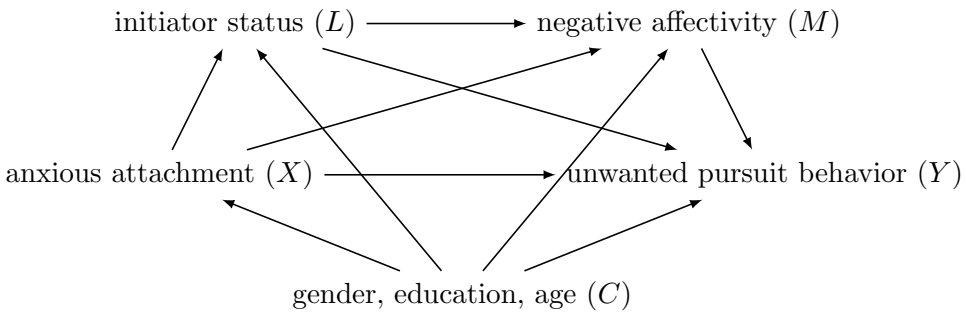

Figure 5: Causal diagram reflecting exposure-induced confounding.

mediated effects through each of the mediators one at a time. That is, if the aforementioned ignorability assumptions (A1-A4) hold with respect to the whole set of mediators, natural indirect effects, as defined as causal pathways through single mediators, are identified since assumptions (A1-A4) then imply that the given mediators are independent (Imai and Yamamoto 2013; VanderWeele and Vansteelandt 2013) given exposure and baseline covariates. Moreover, Lange *et al.* (2014) recently proposed a regression-based approach for testing interdependence between mediators and demonstrated how independent intermediate pathways can be assessed in a single natural effect model using the weighting-based approach.

Often, however, mediators are interdependent and sometimes can be thought of as being linked in a sequential causal chain. For instance, not having initiated the break-up could have made respondents more prone to feeling sad, jealous, angry, frustrated or hurt, as reflected in the causal diagram depicted in Figure 5. Under this diagram, initiator status confounds the relation between the mediator and outcome (given that negative affectivity is the mediator of interest), while at the same time being influenced by the exposure, hence violating identification assumption (A4). As a consequence, the natural indirect effect via negative affectivity can no longer identified using the mediation formula under the causal diagram depicted in Figure 5.

If indications of exposure-induced confounding or mediators affecting one another are present, an alternative might be to consider these multiple mediators jointly and to redefine natural indirect and direct effects by decomposing the total causal effect into an effect mediated through all given mediators simultaneously and an effect not mediated by any of these mediators, respectively (VanderWeele and Vansteelandt 2013; VanderWeele, Vansteelandt, and Robins 2014). Although this type of effect decomposition might not target the initial mediation hypothesis, it may, in certain cases, still shed some light on the underlying causal mechanisms. In particular, it can be interesting to assess if the two mediators in combination leads to a null direct effect as this signals that all important components of the causal chain from exposure to outcome have been identified.

For example, in the natural effect model framework, $\exp(\beta_1^{\star\star})$ in model

$$\text{logit}\Pr\{Y(x, L(x^*), M(x^*)) = 1|C\} = \beta_0^{\star\star} + \beta_1^{\star\star}x + \beta_2^{\star\star}x^* + \beta_3^{\star\star}C, \tag{6}$$

captures the (newly-defined) natural direct effect odds ratio

$$\text{OR}_{1,0|C}^{\text{NDE}} = \frac{\text{odds}\{Y(1, L(x^*), M(x^*)) = 1|C\}}{\text{odds}\{Y(0, L(x^*), M(x^*)) = 1|C\}},$$

whereas $\exp(\beta_2^{\star\star})$ captures the natural indirect effect odds ratio

$$\mathrm{OR}_{1,0|C}^{\mathrm{NIE}} = \frac{\mathrm{odds}\left\{Y(x, L(1), M(1)) = 1|C\right\}}{\mathrm{odds}\left\{Y(x, L(0), M(0)) = 1|C\right\}}$$

through L and M jointly.

Fitting this natural effect model, however, requires both mediators to be taken into account in the working model(s). When applying the weighting-based approach, dealing with multiple mediators entails fitting a model for each of the mediators separately to calculate ratio-of-mediator probability weights, as in Lange *et al.* (2014). The imputation-based approach, on the other hand, only requires one working model for the outcome (i.e., an imputation model). For this reason, estimation of joint mediated effects is implemented only for the imputation-based approach in the current version of the **medflex** package.

Hence, after expanding the data and imputing counterfactual outcomes by fitted values based on an imputation model conditional on both L and M, for instance the logistic model

$$\mathrm{logit}\,\Pr(Y = 1|X, L, M, C) = \gamma_0^{\star\star} + \gamma_1^{\star\star}X + \gamma_2^{\star\star}L + \gamma_3^{\star\star}M + \gamma_4^{\star\star}L \cdot M + \gamma_5^{\star\star}C,$$

which also allows for a mediator-mediator interaction, one can fit natural effect model (6) to the imputed dataset. In R, these steps can be implemented using the following code:

```
R> impData <- neImpute(UPB ~ att + initiator * negaff + gender + educ + age,
+                       family = binomial("logit"), nMed = 2, data = UPBdata)
R> neMod6 <- neModel(UPB ~ att0 + att1 + gender + educ + age,
+                    family = binomial("logit"), expData = impData,
+                    se = "robust")
R> summary(neMod6)

Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: att
Mediator(s): initiator, negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.03066    0.82004   -1.26  0.20881
att0         0.42728    0.12484    3.42  0.00062 ***
att1         0.26958    0.06467    4.17  3.1e-05 ***
genderM      0.15883    0.28656    0.55  0.57940
educM       -0.29465    0.54294   -0.54  0.58734
educH       -0.40742    0.55354   -0.74  0.46171
age         -0.00838    0.01545   -0.54  0.58758
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As illustrated in the R code, the number of mediators to be considered jointly should be set via the `nMed` argument in the `neImpute` function. If `nMed = 2`, not only the second predictor

variable, but the two predictor variables declared after the exposure variable are internally coded as mediators. Correct specification of the (number of) mediators can easily be checked in the `summary` output of the natural effect model object, which mentions the name of the exposure and all mediators.

Although we have hypothesized that initiator status affects the level of experienced negative affectivity, this joint mediator approach does not necessarily require knowing the ordering of the mediators. VanderWeele and Vansteelandt (2013) and VanderWeele *et al.* (2014) described how additional insight into the causal mechanisms can be gained when the ordering is (assumed to be) known. These authors advocated a sequential approach which enables further effect decomposition of the total causal effect into multiple path-specific effects (Avin *et al.* 2005). Such sequential approach can easily be embedded in the natural effect model framework and is planned to be implemented in an upcoming version of the **medflex** package.

9. Concluding remarks

In this paper, we provided some theoretical background on the counterfactual framework, in particular on mediation analysis and natural direct and indirect effects, and described the functionalities of the R package **medflex**.

This package combines some important strengths of other (software) applications for mediation analysis that build on the mediation formula, while accommodating some of their respective weaknesses. The major appeal of this package is its flexibility in dealing with non-linear parametric models and the functionalities it offers for hypothesis testing by resorting to natural effect models, which directly parameterize the target causal estimands. Furthermore, for the most common parametric models, robust standard errors can be obtained, so the computer-intensive bootstrap can be avoided. A limitation of this package is that, at present, it does not offer a framework for sensitivity analysis for possible violations of the identification assumptions of the causal estimands.

As mentioned in the previous section, additional functionalities for dealing with exposure-induced confounding and multiple mediators are intended to be added to the package in the future, as well as extensions for survival models. Future developments within the natural effect model framework will be added in updates of the package.

A. Link between estimators and the mediation formula

A.1. Weighting-based estimator (Lange *et al.* 2012)

Fitting a stratum-specific natural effect model using the weighting-based approach requires a model for the mediator distribution $\Pr(M|X,C)$ as a working model:

$$\begin{aligned}
\mathsf{E}\left\{Y(x,M(x^*))|C\right\} &= \sum_m \mathsf{E}(Y|X=x,M=m,C)\Pr(M=m|X=x^*,C) \\
&= \sum_y \sum_m y \cdot \Pr(Y=y|X=x,M=m,C)\Pr(M=m|X=x^*,C) \\
&= \sum_y \sum_m y \cdot \frac{\Pr(Y=y,M=m|X=x,C)}{\Pr(M=m|X=x,C)}\Pr(M=m|X=x^*,C) \\
&= \mathsf{E}\left[Y \cdot \frac{\Pr(M=m|X=x^*,C)}{\Pr(M=m|X=x,C)} \;\middle|\; X=x,C \right]
\end{aligned}$$

A.2. Imputation-based estimator (Vansteelandt *et al.* 2012)

Fitting a stratum-specific natural effect model using the imputation-based approach requires an imputation model for the mean outcome $\mathsf{E}(Y|X,M,C)$ as a working model:

$$\begin{aligned}
\mathsf{E}\left\{Y(x,M(x^*))|C\right\} &= \sum_m \mathsf{E}(Y|X=x,M=m,C)\Pr(M=m|X=x^*,C) \\
&= \mathsf{E}\left[\mathsf{E}(Y|X=x,M,C) \;\middle|\; X=x^*,C \right] \qquad (7)
\end{aligned}$$

A.3. Weighted weighting-based estimator (Lange *et al.* 2012)

Fitting a marginal or population-averaged natural effect model requires a propensity score model for the exposure $\Pr(X|C)$ as additional working model:

$$\begin{aligned}
\mathsf{E}\{Y(x,M(x^*))\} &= \sum_c \sum_m \mathsf{E}(Y|X=x,M=m,C=c)\Pr(M=m|X=x^*,C=c)\Pr(C=c) \\
&= \sum_y \sum_c \sum_m y \cdot \Pr(Y=y|X=x,M=m,C=c) \\
&\qquad \cdot \Pr(M=m|X=x^*,C=c)\frac{\Pr(C=c,X=x)}{\Pr(X=x|C=c)} \\
&= \sum_y \sum_c \sum_m y \cdot \frac{\Pr(Y=y,M=m|X=x,C=c)}{\Pr(M=m|X=x,C=c)} \\
&\qquad \cdot \Pr(M=m|X=x^*,C=c)\frac{\Pr(C=c,X=x)}{\Pr(X=x|C=c)} \\
&= \sum_y \sum_c \sum_m y \cdot \frac{\Pr(Y=y,M=m,C=c,X=x)}{\Pr(X=x|C=c)}\frac{\Pr(M=m|X=x^*,C=c)}{\Pr(M=m|X=x,C=c)}
\end{aligned}$$

$$= \sum_y \sum_c \sum_m y \cdot \frac{\Pr(Y=y, M=m, C=c|X=x)}{\Pr(X=x|C=c)} \Pr(X=x) \cdot \frac{\Pr(M=m|X=x^*, C=c)}{\Pr(M=m|X=x, C=c)}$$

$$= \mathsf{E}\left[\frac{Y}{\Pr(X=x|C)} \frac{\Pr(M|X=x^*, C)}{\Pr(M|X=x, C)} \middle| X=x \right] \cdot \Pr(X=x)$$

$$= \mathsf{E}\left[\frac{YI(X=x)}{\Pr(X=x|C)} \frac{\Pr(M|X=x^*, C)}{\Pr(M|X=x, C)} \right]$$

A.4. Weighted imputation-based estimator (related to Albert 2012)

$$\mathsf{E}\{Y(x, M(x^*))\} = \sum_c \sum_m \mathsf{E}(Y|X=x, M=m, C=c) \Pr(M=m|X=x^*, C=c) \Pr(C=c)$$

$$= \sum_c \sum_m \frac{\mathsf{E}(Y|X=x, M=m, C=c)}{\Pr(X=x^*|C=c)} \Pr(M=m, C=c, X=x^*)$$

$$= \sum_c \sum_m \frac{\mathsf{E}(Y|X=x, M=m, C=c)}{\Pr(X=x^*|C=c)} \Pr(M=m, C=c|X=x^*) \Pr(X=x^*)$$

$$= \mathsf{E}\left[\frac{\mathsf{E}(Y|X=x, M, C)}{\Pr(X=x^*|C)} \middle| X=x^* \right] \cdot \Pr(X=x^*)$$

$$= \mathsf{E}\left[\frac{\mathsf{E}(Y|X=x, M, C)}{\Pr(X=x^*|C)} I(X=x^*) \right]$$

References

Albert JM (2008). "Mediation analysis via potential outcomes models." *Statistics in Medicine*, **27**, 1282–1304.

Albert JM (2012). "Distribution-free mediation analysis for nonlinear models with confounding." *Epidemiology*, **23**(6), 879–88.

Albert JM, Nelson S (2011). "Generalized causal mediation analysis." *Biometrics*, **67**(3), 1028–38.

Avin C, Shpitser I, Pearl J (2005). "Identifiability of Path-specific Effects." In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pp. 357–363. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. URL http://ijcai.org/PastProceedings/IJCAI-05/PDF/0886.pdf.

Baron RM, Kenny DA (1986). "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of Personality and Social Psychology*, **51**(6), 1173–1182.

Bullock JG, Green DP, Ha SE (2010). "Yes, but what's the mechanism? (don't expect an easy answer)." *Journal of Personality and Social Psychology*, **98**(4), 550–558.

Canty A, Ripley BD (2014). ***boot***: *Bootstrap R (S-Plus) Functions*. R package version 1.3-13.

Daniel RM, De Stavola BL, Cousens SN (2011). "gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula." *Stata Journal*, **11**(4), 479–517.

De Smet O, Loeys T, Buysse A (2012). "Post-Breakup Unwanted Pursuit: A Refined Analysis of the Role of Romantic Relationship Characteristics." *Journal of Family Violence*, **27**(5), 437–452.

Emsley R, Liu H (2013). "PARAMED: Stata module to perform causal mediation analysis using parametric regression models." URL http://ideas.repec.org/c/boc/bocode/s457581.html.

Fox J, Weisberg S (2011). *An R Companion to Applied Regression*. Second edition. Sage, Thousand Oaks CA. URL http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Ghent University and Catholic University of Louvain (2010). "Interdisciplinary Project for the Optimisation of Separation trajectories - divorce and separation in Flanders." URL http://www.scheidingsonderzoek.ugent.be/index-eng.html.

Hastie T (2013). ***gam***: *Generalized Additive Models*. R package version 1.09.1, URL http://CRAN.R-project.org/package=gam.

Hayes AF, Preacher KJ (2010). "Quantifying and Testing Indirect Effects in Simple Mediation Models When the Constituent Paths Are Nonlinear." *Multivariate Behavioral Research*, **45**(4), 627–660.

Hayes AF, Preacher KJ (2014). "Statistical mediation analysis with a multicategorical independent variable." *The British Journal of Mathematical and Statistical Psychology*, **67**, 451–470.

Hicks R, Tingley D (2011). "Causal mediation analysis." *The Stata Journal*, **11**(4), 1–15.

Holland PW (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association*, **81**(396), 945–960.

Hong G (2010). "Ratio of mediator probability weighting for estimating natural direct and indirect effects." In *Proceedings of the American Statistical Association, Biometrics Section*, pp. 2401–2415. American Statistical Association, Alexandria, VA.

Hothorn T, Bretz F, Westfall P (2008). "Simultaneous Inference in General Parametric Models." *Biometrical Journal*, **50**(3), 346–363.

Iacobucci D (2012). "Mediation analysis and categorical variables: The final frontier." *Journal of Consumer Psychology*, **22**(4), 582–594.

IBM Corporation (2013). *IBM SPSS Statistics, Version 22.0*. IBM Corporation, Armonk, NY. URL http://www-01.ibm.com/software/analytics/spss/.

Imai K, Keele L, Tingley D (2010a). "A general approach to causal mediation analysis." *Psychological Methods*, **15**(4), 309–334.

Imai K, Keele L, Yamamoto T (2010b). "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science*, **25**(1), 51–71.

Imai K, Yamamoto T (2013). "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments." *Political Analysis*, **21**(2), 141–171.

Judd CM, Kenny DA (1981). "Process Analysis: Estimating Mediation in Treatment Evaluations." *Evaluation Review*, **5**(5), 602–619.

Lange T, Rasmussen M, Thygesen LC (2014). "Assessing natural direct and indirect effects through multiple pathways." *American Journal of Epidemiology*, **179**(4), 513–8.

Lange T, Vansteelandt S, Bekaert M (2012). "A Simple Unified Approach for Estimating Natural Direct and Indirect Effects." *American Journal of Epidemiology*, **176**(3), 190–195.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**(1), 13–22.

Loeys T, Moerkerke B, De Smet O, Buysse A, Steen J, Vansteelandt S (2013). "Flexible Mediation Analysis in the Presence of Nonlinear Relations: Beyond the Mediation Formula." *Multivariate Behavioral Research*, **48**(6), 871–894.

MacKinnon DP (2008). *Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, New York. ISBN 9780805864304.

MacKinnon DP, Dwyer JH (1993). "Estimating Mediated Effects in Prevention Studies." *Evaluation Review*, **17**(2), 144–158.

MacKinnon DP, Lockwood CM, Brown CH, Wang W, Hoffman JM (2007). "The intermediate endpoint effect in logistic and probit regression." *Clinical Trials*, **4**, 499–513.

Muller D, Judd CM, Yzerbyt VY (2005). "When moderation is mediated and mediation is moderated." *Journal of Personality and Social Psychology*, **89**(6), 852–63.

Muthén BO, Asparouhov T (2015). "Causal Effects in Mediation Modeling: An Introduction with Applications to Latent Variables." *Structural Equation Modeling*, **22**(1), 12–23.

Muthén LK, Muthén BO (1998-2012). ***Mplus*** *User's Guide. Seventh Edition.* Muthén & Muthén, Los Angeles, CA.

Pearl J (2001). "Direct and indirect effects." In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, UAI'01, pp. 411–420. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ISBN 1-55860-800-1. URL http://dl.acm.org/citation.cfm?id=2074022.2074073.

Pearl J (2012). "The mediation formula: A guide to the assessment of causal pathways in nonlinear models." In C Berzuini, P Dawid, L Bernardinelli (eds.), *Causality: Statistical Perspectives and Applications*, October 2011, pp. 151–179. John Wiley and Sons, Chichester, UK.

Polley E, van der Laan M (2014). ***SuperLearner***: *Super Learner Prediction*. R package version 2.0-15, URL http://CRAN.R-project.org/package=SuperLearner.

Preacher KJ, Rucker DD, Hayes AF (2007). "Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions." *Multivariate Behavioral Research*, **42**(1), 185–227.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robins J (1992). "Estimation of the time-dependent accelerated failure time model in the presence of confounding factors." *Biometrika*, **79**(2), 321–334.

Robins JM, Greenland S (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology*, **3**(2), 143–155.

SAS Institute Inc (2014). *SAS/STAT 13.2*. SAS Institute Inc., Cary, NC. URL http://www.sas.com/.

StataCorp (2013). *Stata Statistical Software: Release 13*. StataCorp LP, College Station, TX. URL http://www.stata.com/.

Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014). "mediation: R Package for Causal Mediation Analysis." *Journal of Statistical Software*, **59**(5).

Valeri L, VanderWeele TJ (2013). "Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros." *Psychological Methods*, **18**(2), 137–50.

van der Laan MJ, Petersen ML (2008). "Direct Effect Models." *The International Journal of Biostatistics*, **4**(1), 1–27.

VanderWeele TJ (2011). "Causal mediation analysis with survival data." *Epidemiology*, **22**(4), 582–585.

VanderWeele TJ (2013). "A Three-way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects." *Epidemiology*, **24**(2), 224–232.

VanderWeele TJ, Vansteelandt S (2009). "Conceptual issues concerning mediation, interventions and composition." *Statistics and Its Interface*, **2**(4), 457–468.

VanderWeele TJ, Vansteelandt S (2010). "Odds ratios for mediation analysis for a dichotomous outcome." *American Journal of Epidemiology*, **172**(12), 1339–48.

VanderWeele TJ, Vansteelandt S (2013). "Mediation Analysis with Multiple Mediators." *Epidemiological Methods*, **2**(1), 95–115.

VanderWeele TJ, Vansteelandt S, Robins JM (2014). "Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder." *Epidemiology*, **25**(2), 300–306.

Vansteelandt S (2012). "Understanding counterfactual-based mediation analysis approaches and their differences." *Epidemiology*, **23**(6), 889–91.

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

Yee TW, Wild CJ (1996). "Vector Generalized Additive Models." *Journal of Royal Statistical Society, Series B*, **58**(3), 481–493.

Affiliation:

Johan Steen
Department of Applied Mathematics, Computer Science and Statistics
Faculty of Sciences
Ghent University
Krijgslaan 281, S9
9000 Ghent, Belgium
E-mail: johan.steen@ugent.be
URL: http://users.ugent.be/~jsteen/

Tom Loeys
Department of Data Analysis
Faculty of Psychology and Educational Sciences
Ghent University
Henri Dunantlaan 1
9000 Ghent, Belgium
E-mail: tom.loeys@ugent.be

Beatrijs Moerkerke
Department of Data Analysis
Faculty of Psychology and Educational Sciences
Ghent University
Henri Dunantlaan 1
9000 Ghent, Belgium
E-mail: beatrijs.moerkerke@ugent.be

Stijn Vansteelandt
Department of Applied Mathematics, Computer Science and Statistics
Faculty of Sciences
Ghent University
Krijgslaan 281, S9
9000 Ghent, Belgium
E-mail: stijn.vansteelandt@ugent.be