Journal of Statistical Software

February 2017, Volume 76, Issue 11. doi: 10.18637/jss.v076.i11

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 inter-

mediate 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 developments and improved both inference and interpretability of direct and indirect effect estimates in nonlinear settings by building on the central notion of counterfactual or potential outcomes. 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 2012), respectively.

In this paper, we introduce **medflex** (Steen, Loeys, Moerkerke, and Vansteelandt 2017), an R package that enables 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 2012b), 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 `https://CRAN.R-project.org/package=medflex` (R Core Team 2016).

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 correspond-

ing 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; Petersen, Sinisi, and van der Laan 2006; Imai *et al.* 2010b), 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 focus 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 (i.e., exposure-mediator interactions and moderated mediation) (Section 5). Tools are provided for calculating 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). In Section 9, we further elaborate on modeling demands and missing data, two aspects that may need to be taken into consideration by practitioners when choosing between the two main estimation approaches offered by the package. Finally, we conclude with some final remarks and list some extensions of the package which are planned to be implemented in the near future (Section 10).

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. Readers familiar with counterfactual notation, definitions and assumptions for natural direct and indirect effects may wish to skip to Section 2.2.

Let $Y_i(x)$ denote the potential outcome for subject i that had been observed if, possibly contrary to the fact, i had been assigned to treatment (or exposure level) x. For a binary exposure (with $X = 1$ for the exposed and $X = 0$ for the unexposed), the individual-level causal effect can then be expressed by comparing $Y_i(1)$ to $Y_i(0)$, whereas the population average total causal effect can be expressed as $\mathsf{E}\{Y(1) - Y(0)\}$. Similarly, direct and indirect effects have been defined in terms of counterfactual outcomes. For instance, the definition of the so-called *controlled* direct effect reflects the traditional notion of measuring the effect of exposure while fixing the mediator M at the same value m for all subjects (Robins and

Greenland 1992). Using counterfactual notation, this effect can be expressed as

$$\text{CDE}(m) = \mathsf{E}\{Y(1, m) - Y(0, m)\},$$

where $Y(x, m)$ denotes the potential outcome that would have been observed under exposure level x and mediator value m.

Robins and Greenland (1992) introduced an alternative definition that invokes so-called *composite* or *nested* counterfactuals, $Y(x, M(x^*))$. For instance, the (pure) natural direct effect

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

expresses the expected exposure-induced change in outcome when keeping the mediator fixed at the value that had naturally been observed if unexposed. By considering potential intermediate outcomes $M(x^*)$ rather than a fixed mediator value m, these authors offered a definition of direct effect that both allows for natural variation in the mediator and provides a complementary operational definition for the indirect effect (which the definition of the controlled direct effect does not). That is, under the composition assumption, which states that $Y(x, M(x)) = Y(x)$ (VanderWeele and Vansteelandt 2009), the difference between the average total effect $\mathsf{E}\{Y(1) - Y(0)\}$ and the (pure) natural direct effect yields an expression for the (total) natural indirect effect

$$\text{NIE}(1) = \mathsf{E}\{Y(1, M(1)) - Y(1, M(0))\}.$$

This reflects the expected difference in outcome if all subjects were exposed but their mediator value had changed to the value it would take if unexposed.

Adopting this counterfactual notation naturally leads to framing causal inference as a missing data problem (Holland 1986): for each subject i, only one counterfactual outcome, i.e., $Y_i = Y_i(X_i, M_i(X_i))$, is observed. Consequently, identification of natural effects relies on rather strong causal assumptions. In the context of mediation analysis, the most commonly invoked conditions for identification can be encoded in a causal diagram (such as Figure 2) interpreted as a non-parametric structural equation model with independent error terms (NPSEM-IE; Pearl 2001). More specifically, upon adjustment for a given set of observed baseline covariates C, such model implies certain independencies among variables and potential outcomes (A1–A4) which have been proposed as sufficient conditions for non-parametric or model-free identification of natural effects. However, this adjustment set C needs to be carefully selected[1], such that it is deemed sufficient to control for confounding (i) between exposure and outcome, thereby satisfying

$$Y(x, m) \perp\!\!\!\perp X | C \qquad \text{for all levels of } x \text{ and } m, \tag{A1}$$

[1] Pearl (2001, 2014) offers a weaker set of conditions, which does not require a common set of baseline covariates to deconfound each of the possibly confounded relations, but allows for separate adjustment sets for the exposure-mediator relation on the one hand and the exposure-outcome and mediator-outcome relations on the other hand. This set of conditions is considered weaker, since it allows for identification under certain non-parametric structural equation models with unobserved confounders. Although, theoretically, the natural effect model framework can be shown to easily accommodate separate adjustment sets, this has not been implemented as such in the **medflex** package. However, as Imai, Keele, Tingley, and Yamamoto (2014) argue, this weaker set of conditions might be of little practical relevance since researchers in most settings lack sufficient knowledge about the precise structure of confounding. Nonetheless, the estimation algorithms implemented in the **mediation** package (Tingley *et al.* 2014), easily allow to specify separate adjustment sets (Imai *et al.* 2014).

(ii) between exposure and mediator, thereby satisfying

$$M(x) \perp\!\!\!\perp X|C \qquad \text{for all levels of } x, \tag{A2}$$

and (iii) between mediator and outcome (after adjustment for the exposure), thereby satisfying

$$Y(x,m) \perp\!\!\!\perp M|X=x,C \qquad \text{for all levels of } x \text{ and } m. \tag{A3}$$

In addition to these 'no omitted variables' assumptions (A1–A3), identification requires the further 'cross-worlds independence' assumption (Pearl 2001)

$$Y(x,m) \perp\!\!\!\perp M(x^*)|C \qquad \text{for all levels of } x,\ x^* \text{ and } m, \tag{A4}$$

which is satisfied under a NPSEM-IE when no confounders of the mediator-outcome relationship (whether observed or unobserved) are affected by the exposure (i.e., no intermediate or exposure-induced confounding).

Whereas the first two assumptions by definition hold in randomized experiments, the other two assumptions may not.[2] Although Judd and Kenny (1981) initially pointed to its importance, condition A3 since has largely been ignored in much of the social sciences literature, as evidenced by many mediation studies not adjusting for confounders of the mediator-outcome 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; Mayer, Thoemmes, Rose, Steyer, and West 2014).

Condition A4 is more difficult to grasp intuitively. It is a strong assumption because, in contrast to the other three conditions, it is impossible to design a study that would be able to validate it (Robins and Richardson 2010; although see Imai, Tingley, and Yamamoto 2013 for a notable attempt).

The interested reader can refer to VanderWeele and Vansteelandt (2009) for a more detailed and intuitive account of these identification assumptions (or to Petersen *et al.* 2006 or Imai *et al.* 2010b for alternative sets of assumptions).

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 parametric linear regression. The main identification result (Pearl 2001; Imai *et al.* 2010b), which Pearl (2012) referred to as the *mediation formula*, has played a pivotal role in this regard. 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

[2]Note that A1 is sufficient for identifying total causal effects, whereas identification of controlled direct effects can be obtained under assumptions A1 and A3.

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).[3]

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 additionally allows to specify semi-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.

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.* 2012b). For instance, when using logistic regression models

$$\begin{aligned} \operatorname{logit} \Pr(M = 1|X, C) &= \alpha_0 + \alpha_1 X + \alpha_2 C \\ \operatorname{logit} \Pr(Y = 1|X, M, C) &= \beta_0 + \beta_1 X + \beta_2 M + \beta_3 C \end{aligned} \tag{0}$$

for binary mediators and outcomes, the mediation formula yields

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

[3] 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.* 2012b).

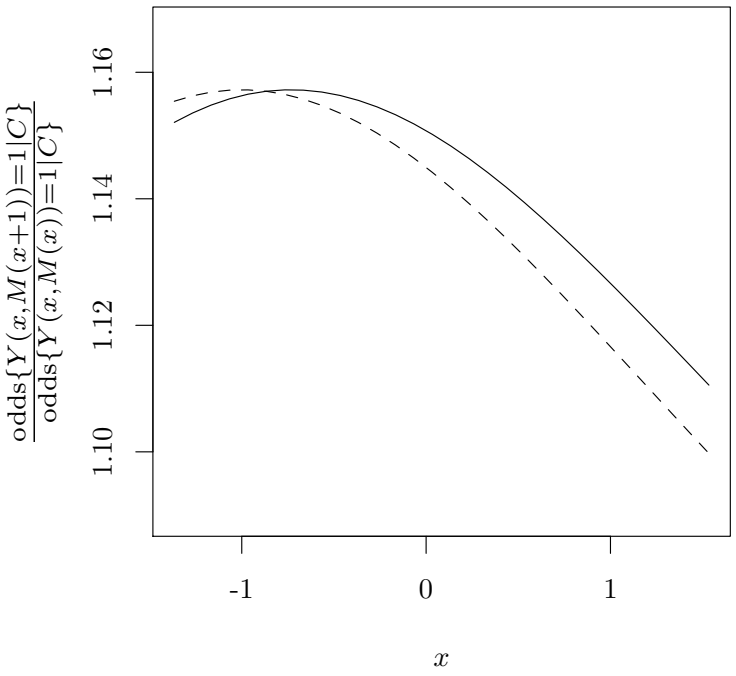

Figure 1: Estimated (total) natural indirect effect odds ratios corresponding to a one-unit change in anxious attachment level as a function of different reference levels for anxious attachment level x (as obtained through direct application of the mediation formula). These are conditional estimates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

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, the corresponding direct and indirect effect 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.* (2012b), these convoluted expressions render results difficult to report and hypothesis testing (e.g., testing for moderated mediation) infeasible, as it may turn out impossible to find plausible models for the mediator and outcome that combine into effect expressions that do not depend on covariate levels. 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 inter-

actions or moderated mediation), these hypothesis tests probe research questions in terms of e.g., 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 recently proposed by Lange *et al.* (2012) and Vansteelandt *et al.* (2012b). 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),
- estimates can be expressed on more natural effect scales (i.e., a scale that corresponds to the link-function of the outcome model), thereby avoiding potential induced dependence on exposure or covariate levels characteristic for the additive scale,
- natural effect models simplify testing since the hypotheses of interest can always be captured by a finite set of 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.* (2012b).

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. 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 more natural, multiplicative scale. Specifically, in this model, $\exp(\beta_1)$ captures the natural direct effect rate ratio

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

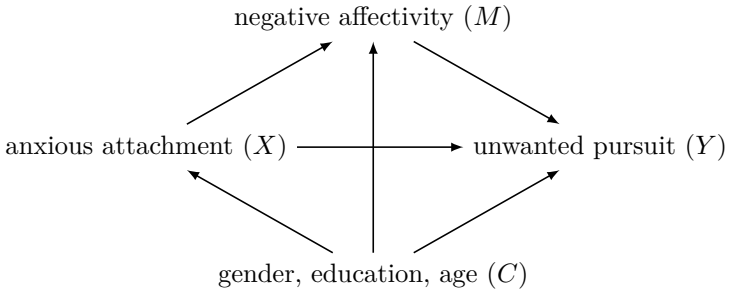

Figure 2: Causal diagram reflecting the mediation hypothesis.

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, this facilitates interpretation and hypothesis testing in nonlinear settings.

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   1
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
```

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 (Pearl 1995) that reflects this mediation hypothesis along with its aforementioned identification assumptions.

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

Table 1: Schematic display of the expanded dataset with missing counterfactual outcomes.

As direct and indirect effects are most easily understood for a binary exposure, 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), $M(x^*)$ to the level of negative affectivity that would have been reported if anxious attachment level were set to x^*, and $Y(x, M(x^*))$ 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^*. To control for confounding, we condition on a set of baseline covariates C: 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). As emphasized earlier, the selection of such an adjustment set needs careful consideration in order to meet identification conditions A1–A4. For illustrative purposes, the current set of baseline covariates C will, possibly contrary to the fact, be considered sufficient to control for confounding throughout the remainder of the paper.

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, as illustrated by Table 1. Although, for the data at hand, three (x, x^*) combinations are unobserved for each individual, to disentangle natural direct and indirect effects, it is sufficient to introduce only one additional observation corresponding to an unobserved combination for which x does not equal x^*.

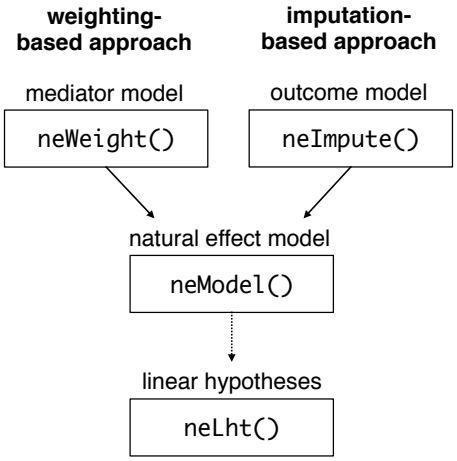

Figure 3: Workflow of the **medflex** package.

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 highly similar 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 sections, we explain both approaches and give example code in R.

3.2. Weighting-based approach

One way to account for missingness in the expanded data is to standardize observed outcomes to the mediator distribution of the hypothetical exposure level x^*. Building on Hong's (2010) ratio-of-mediator-probability weighting (RMPW) method, Lange *et al.* (2012) proposed to weight each observation in the expanded dataset by

$$w_i = \frac{p_i(x^*)}{p_i(x)} = \frac{\Pr(M = M_i | X = x^*, C = C_i)}{\Pr(M = M_i | X = x, C = C_i)}.$$

For instance, for a binary exposure, $\mathsf{E}\{Y(0, M(0))|C\}$ and $\mathsf{E}\{Y(1, M(1))|C\}$ can readily be estimated from the observed data (under assumption A1) without weighting (i.e., as $x = x^*$ the corresponding weights equal 1). To enable estimation of $\mathsf{E}\{Y(1, M(0))|C\}$ and $\mathsf{E}\{Y(0, M(1))|C\}$ RMPW aims to construct a 'parallel' pseudo-population for each exposure group x (within each stratum of C) with mediator values that would have been observed if each subject had been a member of the opposite exposure group $x^* = 1 - x$. This is done by up-weighting individuals whose observed mediator value is more typical for the opposite exposure group than the exposure group to which they originally belong. Similarly, individuals whose observed mediator value is relatively more typical for the original exposure group are

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 2: Schematic display of the weighting-based approach.

down-weighted.[4]

Data expansion hence only requires x^* to take on values different from the observed exposure to enable estimation of natural direct and indirect effects via the weighting-based approach, as illustrated in Table 2. Estimates 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. This procedure easily extends to continuous exposures (see Section 4.2) and/or mediators (provided probabilities are replaced by densities). The interested reader is referred to Appendix A.1, where a more technical account is given on the link between the weighting-based approach and the mediation formula.

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 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:

[4]Hong, Deutsch, and Hill (2015) gives a more detailed example which may provide more intuition into RMPW. Other weighting methods based on inverse odds (ratio) weighting (Huber 2013; Tchetgen Tchetgen 2013) have been proposed recently. In contrast to RMPW these weighting methods rely on models for the exposure distribution (conditional on mediator and baseline covariates). Although these could easily be adopted within the natural effect model framework, these are currently not implemented in the **medflex** package.

```
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`.

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

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).[5]

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)

 [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 bootstrapped standard errors. In order to approximate

[5]In the current version of the package also `vglm` and `vgam` from the **VGAM** package and `gam` from the **gam** package (Hastie 2016) 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.

the sampling distribution of each of the natural effect model parameters, the applied non-parametric bootstrap procedure repeatedly resamples the original data with replacement. For each replication, all aforementioned steps are repeated and estimates of the natural effect model parameters are obtained. The resulting bootstrap distribution can then be used for statistical inference. By refitting the same model for the mediator distribution to each bootstrap sample and recalculating ratio-of-mediator-probability weights for the (subsequently) expanded bootstrap samples, uncertainty related to estimation of the mediator model is incorporated into the bootstrapped standard errors. The number of bootstrap replications defaults to 1000 and can be set in the `nBoot` argument:

```
R> set.seed(1234)
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 statistics 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) -0.92521    0.91389   -1.01    0.311
attbin01     0.39592    0.23283    1.70    0.089 .
attbin11     0.35197    0.08829    3.99  6.7e-05 ***
genderM      0.27597    0.23954    1.15    0.249
educM        0.16701    0.75404    0.22    0.825
educH        0.42335    0.75101    0.56    0.573
age         -0.00945    0.01283   -0.74    0.461
---
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) -0.92521    0.71463   -1.29    0.195
attbin01     0.39592    0.21761    1.82    0.069 .
attbin11     0.35197    0.08939    3.94  8.2e-05 ***
genderM      0.27597    0.23370    1.18    0.238
educM        0.16701    0.50065    0.33    0.739
educH        0.42335    0.50917    0.83    0.406
age         -0.00945    0.01227   -0.77    0.441
---
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 at any given level of anxious attachment x, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\mathrm{OR}}^{\mathrm{NDE}}_{1,0|C} = \frac{\mathrm{odds}\left\{Y(1, M(x)) = 1|C\right\}}{\mathrm{odds}\left\{Y(0, M(x)) = 1|C\right\}} = \exp(\hat{\beta}_1) = \exp(0.3959) = 1.49.$$

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

$$\widehat{\mathrm{OR}}^{\mathrm{NIE}}_{1,0|C} = \frac{\mathrm{odds}\left\{Y(x, M(1)) = 1|C\right\}}{\mathrm{odds}\left\{Y(x, M(0)) = 1|C\right\}} = \exp(\hat{\beta}_2) = \exp(0.352) = 1.42.$$

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    0.97    2.28
attbin11    1.19    1.69
```

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

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

(`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.[6]

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.* 2012b). By setting x^* (rather than x) equal to the observed exposure X, unobserved nested counterfactuals can be imputed using any appropriate model for the outcome mean. That is, since the potential intermediate outcome $M(x^*)$ equals the observed mediator M within the subgroup with exposure $X = x^*$, $Y(x, M(x^*))$ equals $Y(x, M)$ for all individuals in that exposure group. 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 X set to x and with mediator M and baseline covariates C set to their observed values. This approach easily accommodates missing outcomes in the original dataset, as the corresponding nested counterfactuals can likewise be imputed.

In contrast to the weighting-based approach, data expansion only requires x to take on values different from the observed exposure to enable estimation of natural direct and indirect effects, as illustrated in Table 3. Estimates can finally be obtained upon fitting a natural effect model to the imputed dataset. For ease of implementation, observed nested counterfactuals are imputed as well in the **medflex** package.[7] In Appendix A.2, we demonstrate the link between the mediation formula and the imputation-based approach by showing how the former can be rewritten as an expression that prescribes estimating nested counterfactuals by calculating the mean of imputed nested counterfactuals, conditional on x, x^* and C.

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.

[6]The `type` argument in `confint` corresponds to that of the `boot.ci` function from the **boot** package (Canty and Ripley 2016), which is called internally.

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

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) (i.e., it should at least contain all terms 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. All other predictors are automatically coded as baseline covariates. It is important to adhere to this prespecified order to enable `neImpute` to create valid pointers to these different types of predictor variables. This requirement extends to the use of operators different from the `+` operator, such as the `:` and `*` operators (when e.g., adding interaction terms). For instance, the formula expressions `Y ~ X + M + C1 + C2 + X:C1 + M:C1`, `Y ~ X + M + X:C1 + M:C1 + C1 + C2`, `Y ~ (X + M) * C1 + C2` and `Y ~ X * C1 + M * C1 + C2` all impose the same structural form for the imputation model. However, only for the former three expressions, correct pointers to exposure, mediator and baseline covariates will be created, as the order of occurrence of each of the unique predictor variables is identical in all three specifications, but not in the latter.

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. The outcomes are no longer binary, but are substituted by conditional mean imputations.

```
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 0.492
2  1       0       1  1.001      M   0.84    myself      F    M  41 0.384
3  2       0       0 -0.709      L  -1.26      both      M    M  42 0.187
4  2       1       0 -0.709      L  -1.26      both      M    M  42 0.263
```

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)  -0.9216     0.6892   -1.34     0.18
attbin01      0.4015     0.2134    1.88     0.06 .
attbin11      0.3407     0.0805    4.23  2.3e-05 ***
genderM       0.2940     0.2250    1.31     0.19
educM         0.3462     0.4817    0.72     0.47
educH         0.5143     0.4878    1.05     0.29
age          -0.0122     0.0119   -1.02     0.31
---
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 sections, we give a detailed description on how to fit natural effect models with multicategorical (i.e., ordinal or nominal) and continuous exposures. In these sections, 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 4. When using the weighting-based approach, models for binary, count and

	Outcome type					
	Binary		Count		Continuous	
Mediator type	`neWeight`	`neImpute`	`neWeight`	`neImpute`	`neWeight`	`neImpute`
Binary	✓	✓	✓	✓	✓	✓
Count	✓	✓	✓	✓	✓	✓
Continuous	✓	✓	✓	✓	✓	✓
Ordinal		✓		✓		✓
Nominal	✓*	✓	✓*	✓	✓*	✓

Table 4: 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.

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`).[8] 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, 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 0.468
```

[8] 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.

```
2  1       H       M  1.001      1   0.84    myself      F    M  41 0.558
3  1       L       M  1.001      1   0.84    myself      F    M  41 0.366
4  2       L       L -0.709      0  -1.26      both      M    M  42 0.182
5  2       M       L -0.709      0  -1.26      both      M    M  42 0.253
6  2       H       L -0.709      0  -1.26      both      M    M  42 0.327
```

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).

```
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)  -0.9616     0.6976   -1.38  0.16807
attcat0M      0.3921     0.2365    1.66  0.09729 .
attcat0H      0.7239     0.3105    2.33  0.01975 *
attcat1M      0.3012     0.0797    3.78  0.00016 ***
attcat1H      0.5218     0.1314    3.97  7.2e-05 ***
genderM       0.2700     0.2266    1.19  0.23336
educM         0.3279     0.4817    0.68  0.49601
educH         0.4826     0.4877    0.99  0.32239
age          -0.0127     0.0121   -1.05  0.29510
---
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  5.98       0.05 .
```

```
attcat1  2 19.11    7.1e-05 ***
gender   1  1.42       0.23
educ     2  1.17       0.56
age      1  1.10       0.30
---
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 exposure levels to construct auxiliary variables x and x^* for each subject, Vansteelandt *et al.* (2012b) proposed to draw specific quantiles from the conditional density 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.[9]

Both `neWeight` and `neImpute` allow to choose the number of draws to sample from this conditional density via the `nRep` argument (which defaults to 5).[10]

```
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 0.309
2  1  8.02e-06  1.001      1      M   0.84    myself      F    M  41 0.429
3  1  1.64e+00  1.001      1      M   0.84    myself      F    M  41 0.557
4  2 -1.66e+00 -0.709      0      L  -1.26      both      M    M  42 0.149
5  2 -1.82e-02 -0.709      0      L  -1.26      both      M    M  42 0.227
6  2  1.63e+00 -0.709      0      L  -1.26      both      M    M  42 0.330
```

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

[9] 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.

[10] We recommend to use a minimum of 3 draws. Although finite sample bias and sampling variability can be reduced to some extent by choosing a larger number of draws, simulations have shown this gain to be ignorable when choosing more than 5 draws (Vansteelandt *et al.* 2012b).

```
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)  -0.4873     0.6862   -0.71   0.4776
att0          0.2923     0.1091    2.68   0.0074 **
att1          0.2018     0.0470    4.29  1.8e-05 ***
genderM       0.2671     0.2274    1.17   0.2402
educM         0.2679     0.4894    0.55   0.5841
educH         0.4103     0.4959    0.83   0.4080
age          -0.0120     0.0122   -0.99   0.3236
---
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

$$\widehat{\text{OR}}^{\text{NDE}}_{x+1,x|C} = \frac{\text{odds}\{Y(x+1, M(x)) = 1|C\}}{\text{odds}\{Y(x, M(x)) = 1|C\}} = \exp(\hat{\beta}_1) = \exp(0.29) = 1.34, \quad \text{and}$$
$$\widehat{\text{OR}}^{\text{NIE}}_{x+1,x|C} = \frac{\text{odds}\{Y(x, M(x+1)) = 1|C\}}{\text{odds}\{Y(x, M(x)) = 1|C\}} = \exp(\hat{\beta}_2) = \exp(0.2) = 1.22,$$

respectively, regardless of the initial level x. Defining natural effects on the risk difference scale (as in **mediation**) 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

So far, the considered natural effect models reflected the assumption that exposure and mediator do not interact in their effect on the outcome (on the scale defined by the link function).

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 postulated to be the same for each choice of mediator level $M(x^*)$, and hence for each choice of reference exposure level x^*, at which the mediator is evaluated. 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 postulated to be constant across different choices of x at which the outcome is 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. However, in many studies, these assumptions may not 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, the total effect odds ratio

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

can be expressed as the product

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

of the pure direct and pure indirect effect odds ratios and the mediated interaction odds ratio. 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, β'_3 in natural effect model

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

differs from 0. When including this interaction term in the outcome model, β'_1 and β'_2 encode 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, a minimal imputation model for natural effect model (2) would be the logistic regression model

$$\mathrm{logit}\,\Pr(Y = 1|X, M, C) = \gamma'_0 + \gamma'_1 X + \gamma'_2 M + \gamma'_3 X M + \gamma'_4 C.$$

The output of the corresponding natural effect model object suggests there is no evidence for mediated interaction at the 5% significance level ($p = 0.0541$).

```
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)  -0.3949     0.6800   -0.58   0.5614
att0          0.2950     0.1102    2.68   0.0074 **
att1          0.1817     0.0467    3.90  9.8e-05 ***
genderM       0.2815     0.2263    1.24   0.2135
educM         0.1798     0.4857    0.37   0.7113
educH         0.3105     0.4929    0.63   0.5287
age          -0.0139     0.0122   -1.14   0.2545
att0:att1     0.0698     0.0363    1.93   0.0541 .
---
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). This moderated 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):

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

The amount of effect modification by gender in this model is then simply captured by β_3''.

```
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)   -0.4731     0.6860   -0.69   0.4904
att0           0.2850     0.1069    2.67   0.0077 **
att1           0.1441     0.0583    2.47   0.0134 *
genderM        0.2591     0.2278    1.14   0.2553
educM          0.2718     0.4903    0.55   0.5793
educH          0.4166     0.4975    0.84   0.4024
age           -0.0123     0.0122   -1.00   0.3153
att1:genderM   0.1598     0.1016    1.57   0.1156
---
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 ($p = 0.1156$).

In a similar way, researchers can gauge 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

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

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 calculating and visualizing causal effect estimates

In this section, we highlight tools that can aid in calculating and visualizing specific causal effect estimates of interest. These tools might prove 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 obtained 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 $\beta_1' + \beta_3'$ and $\beta_2' + \beta_3'$, respectively. Similarly, the total causal effect log odds ratio is captured by $\beta_1' + \beta_2' + \beta_3'$. 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        1.44    1.15    1.80
att1 + att0:att1        1.29    1.15    1.43
att0 + att1 + att0:att1 1.73    1.39    2.15
```

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           0.3648     0.1145    3.19   0.0014 **
att1 + att0:att1           0.2515     0.0553    4.55  5.4e-06 ***
att0 + att1 + att0:att1    0.5465     0.1118    4.89  1.0e-06 ***
---
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 interest is mainly focused on 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      0.2950     0.1102    2.68   0.0074 **
total direct effect     0.3648     0.1145    3.19   0.0014 **
pure indirect effect    0.1817     0.0467    3.90  9.8e-05 ***
total indirect effect   0.2515     0.0553    4.55  5.4e-06 ***
total effect            0.5465     0.1118    4.89  1.0e-06 ***
---
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

$$\widehat{\mathrm{OR}}^{\mathrm{NDE}}_{1,0|C}(0) = \frac{\mathrm{odds}\left\{Y(1, M(0)) = 1|C\right\}}{\mathrm{odds}\left\{Y(0, M(0)) = 1|C\right\}} = \exp(\hat{\beta}'_1) = 1.34$$

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

$$\widehat{\mathrm{OR}}^{\mathrm{NDE}}_{1,0|C}(1) = \frac{\mathrm{odds}\left\{Y(1, M(1)) = 1|C\right\}}{\mathrm{odds}\left\{Y(0, M(1)) = 1|C\right\}} = \exp(\hat{\beta}'_1 + \hat{\beta}'_3) = 1.44$$

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

On the other hand, altering negative affectivity from levels that would have been observed at average levels of anxious attachment to levels that would have been observed at attachment

scores of one standard deviation higher, increases the odds of displaying unwanted pursuit behaviors with a factor

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

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

$$\widehat{\text{OR}}^{\text{NIE}}_{1,0|C}(1) = \frac{\text{odds}\left\{Y(1, M(1)) = 1|C\right\}}{\text{odds}\left\{Y(1, M(0)) = 1|C\right\}} = \exp(\hat{\beta}'_2 + \hat{\beta}'_3) = 1.29$$

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.

$$\widehat{\text{OR}}_{1,0|C} = \frac{\text{odds}\left\{Y(1, M(1)) = 1|C\right\}}{\text{odds}\left\{Y(0, M(0)) = 1|C\right\}} = \exp(\hat{\beta}'_1 + \hat{\beta}'_2 + \hat{\beta}'_3) = 1.73.$$

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      0.285
natural indirect effect    0.144
total effect               0.429

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      0.285
natural indirect effect    0.304
total effect               0.589

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

6.3. Global hypothesis tests

Wald tests considering all specified linear hypotheses jointly 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 the multicategorical baseline covariate education level via 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   5.2  2     0.0742
```

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.

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 additionally allows to estimate population-average natural effects. As demonstrated in Appendix A.3 and A.4,

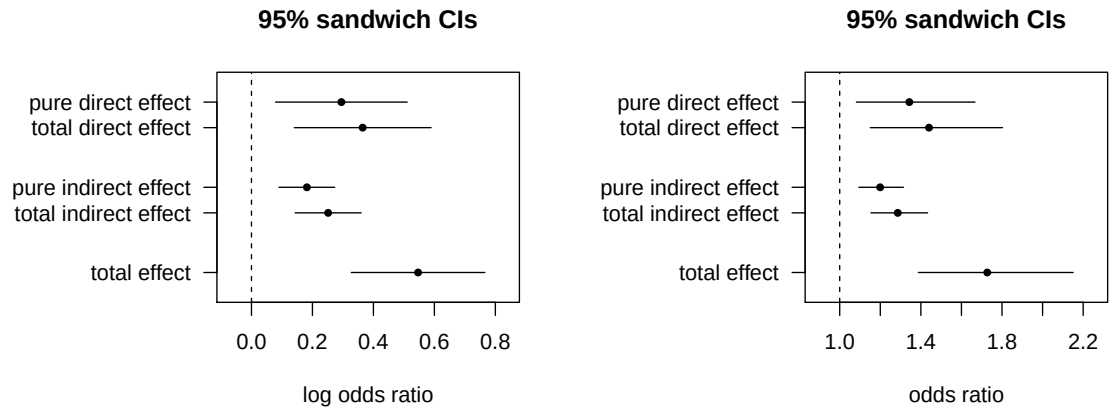

Figure 4: Effect decomposition on the log odds ratio and odds ratio scales.

rewriting the mediation formula reveals that estimation of these population-average effects requires weighting by the reciprocal of the conditional exposure density in order to adjust for confounding (also see Albert 2012; Vansteelandt 2012).

As a consequence, a model for the exposure density 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

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

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)  -0.5793     0.1112   -5.21  1.9e-07 ***
att0          0.2967     0.1082    2.74   0.0061 **
att1          0.2294     0.0578    3.97  7.2e-05 ***
---
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 corresponding to anxious attachment level x^*, increases the odds of displaying unwanted pursuit behaviors with a factor

$$\widehat{\mathrm{OR}}_{1,0}^{\mathrm{NDE}} = \frac{\mathrm{odds}\left\{Y(1, M(x^*)) = 1\right\}}{\mathrm{odds}\left\{Y(0, M(x^*)) = 1\right\}} = \exp(\hat{\theta}_1) = 1.35.$$

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

8. Intermediate confounding: A joint mediation 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. Initiator status (`initiator`: either `"both"`, `"ex-partner"`, or `"myself"`) can thus also be considered a mediator, which we denote L.

If hypothesized mediators are conditionally independent (given exposure and baseline covariates), separate natural effect models can be fitted (each with a different working model involving only one of the mediators) to assess the mediated effects through each of the mediators one at a time. Specifically, if the aforementioned ignorability conditions A1–A4 hold with respect to each mediator separately[11], natural indirect effects, as defined as causal pathways through single mediators, are identified since these conditions imply that the given mediators are independent given exposure and baseline covariates (Imai and Yamamoto 2013; VanderWeele and Vansteelandt 2013). Recently, Lange *et al.* (2014) demonstrated how independent intermediate pathways can be assessed in a single natural effect model using the weighting-based approach. Additionally, these authors proposed a regression-based approach for testing conditional dependence between mediators (also see Loeys *et al.* 2013; Imai and Yamamoto 2013).

Often, however, mediators are interdependent and can be thought of as being linked in a sequential causal chain. For instance, not having initiated the break-up could have made

[11] In addition to the assumptions expressed in A1–A4, we additionally assume that $Y(x,l) \perp\!\!\!\perp X|C$ (for all levels of x and l), $L(x) \perp\!\!\!\perp X|C$ (for all levels of x), $Y(x,l) \perp\!\!\!\perp L|X = x, C$ (for all levels of x and l) and $Y(x,l) \perp\!\!\!\perp L(x^*)|C$ (for all levels of x, x^* and l).

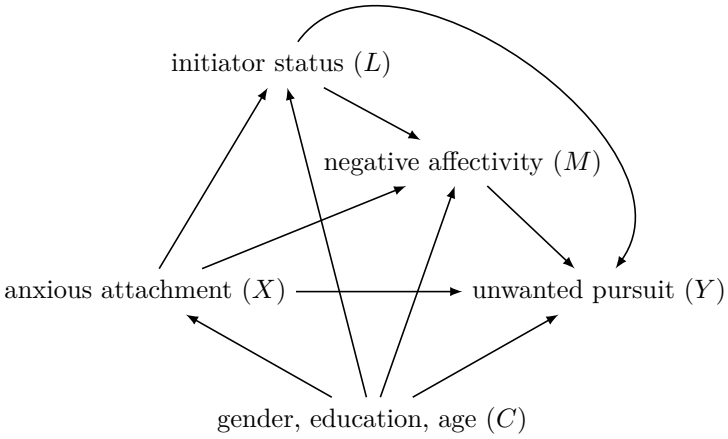

Figure 5: Causal diagram reflecting exposure-induced confounding.

respondents more prone to feeling sad, jealous, angry, frustrated or hurt, as reflected in the causal diagram of 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 affected by the exposure, hence violating identification assumption A4. As a consequence, the natural indirect effect via negative affectivity is no longer identified under the NPSEM-IE depicted in Figure 5 (although see Robins 2003; Tchetgen Tchetgen and VanderWeele 2014; Vansteelandt and VanderWeele 2012, for additional (parametric) restrictions which enable identification). This non-identification can intuitively be appreciated by the fact that, in the presence of an intermediate confounder L, the natural indirect effect via M can be rewritten as

$$\frac{\text{odds}\left\{Y(x, L(x), M(1, L(1))) = 1 | C\right\}}{\text{odds}\left\{Y(x, L(x), M(0, L(0))) = 1 | C\right\}},$$

which involves blocking the causal path through L only ($X \rightarrow L \rightarrow Y$), while at the same time assessing the effect transmitted through L and M ($X \rightarrow L \rightarrow M \rightarrow Y$) (Didelez, Dawid, and Geneletti 2006).

Alternatively, the total causal effect can be decomposed into the effect transmitted through L and M simultaneously and the effect not mediated by any of the given mediators (VanderWeele and Vansteelandt 2013; VanderWeele, Vansteelandt, and Robins 2014). Although such a joint mediation approach might not target the initial mediation hypothesis, it may still shed some light on the underlying causal mechanisms if there are reasons (either theoretical or empirical) to question the validity of condition A4 (with respect to a single mediator)[12], since this decomposition relies on a weaker set of ignorability assumptions. Specifically, if, as under the NPSEM-IE depicted in Figure 5, we assume that a set of baseline covariates C satisfies 'no omitted variables' assumptions A1–A3 with respect to L and M jointly (rather than separately) and that no measured or unmeasured confounders of the $(L, M) - Y$ relation are affected by the exposure[13], the joint mediated effect and corresponding direct effect are

[12]In particular, it can be interesting to assess if the two mediators in combination lead to a null direct effect as this may signal that all important components in the causal chain from exposure to outcome have been identified.

[13]I.e., assuming that $Y(x, l, m) \perp\!\!\!\perp X | C$ (for all levels of x, l and m), $\{M(x), L(x)\} \perp\!\!\!\perp X | C$ (for all levels of x),

identified. The appeal of this joint mediation approach is that by defining a natural indirect effect with respect to a set or vector of mediators (rather than a single mediator) assumption A4 can be made more plausible by simply including mediator-outcome confounders that are deemed likely to be affected by the exposure in the joint set of mediators (VanderWeele and Vansteelandt 2013).

For example, $\exp(\beta_1^{\star\star})$ in model (6)

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

captures the (newly defined) natural direct effect odds ratio

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

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

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

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, is less demanding as it only requires one working model for the outcome. 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, nested counterfactual outcomes need to be imputed by fitted values from an imputation model conditional on both L and M. For instance, in the R code below, a logistic model

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

is fitted that allows the mediators to interact in their effect on the outcome.

```
R> impData <- neImpute(UPB ~ att + initiator * negaff + gender + educ + age,
+    family = binomial("logit"), nMed = 2, data = UPBdata)
```

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. Subsequently, natural effect model (6) can be fitted to the imputed dataset using the `neModel` function.

```
R> neMod6 <- neModel(UPB ~ att0 + att1 + gender + educ + age,
+    family = binomial("logit"), expData = impData, se = "robust")
R> summary(neMod6)
```

$Y(x, l, m) \perp\!\!\!\perp \{L, M\}|X = x, C$ (for all levels of x, l and m) and $Y(x, l, m) \perp\!\!\!\perp \{L(x^*), M(x^*)\}|C$ (for all levels of x, x^*, l and m).

```
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)  -0.4919     0.6854   -0.72    0.473
att0          0.2444     0.1114    2.19    0.028 *
att1          0.2476     0.0538    4.60  4.2e-06 ***
genderM       0.2629     0.2274    1.16    0.248
educM         0.2780     0.4912    0.57    0.571
educH         0.4223     0.4979    0.85    0.396
age          -0.0121     0.0122   -0.99    0.320
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correct specification of the (number of) mediators can easily be checked in the `summary` output of the natural effect model object, which lists the names 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; also see Huber 2013 for an inverse-probability weighting approach and Albert and Nelson 2011 and Daniel, De Stavola, Cousens, and Vansteelandt 2015 for a parametric g-computation approach for estimating some of these path-specific effects). 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. Weighting or imputing?

For both the weighting- and imputation-based approach, valid estimation of natural effects hinges on adequate specification of their corresponding nuisance working models and the natural effect model. In this section, we highlight the impact of model misspecification for each of the two proposed estimation approaches. The resulting trade-off in terms of modeling demands may serve as a guideline as to which of the two approaches is to be preferred in which particular setting. Moreover, certain missing data patterns might also favor one approach over the other, as discussed in more detail below.

9.1. Modeling demands

The proposed weighting-based approach yields consistent natural effects estimates if both the natural effect model and the conditional distribution of the mediator are correctly specified.

The latter needs careful consideration, especially when exposure or baseline covariates are highly predictive of the mediator, for then even minor misspecifications in its conditional expectation can have a major impact on the weights and lead to heavily biased estimation of the target natural effects parameters. However, residual plots with scatterplot smoothers are often helpful to diagnose model inadequacy and can be requested, for instance, by passing the `expData`-class object to the `residualPlots` function from the **car** package. When dealing with continuous mediators, correct modeling not only demands adequate specification of the mediator's expectation, but also requires additional parametric assumptions on the mediator's conditional density (i.e., the distribution of the error terms).[14] Moreover, even under proper model specification, weights for continuous mediators typically tend to be unstable, leading to less precise natural effect estimates and considerable finite sample bias. In particular, when the outcome is linear in the mediator, it might be sensible to avoid unnecessary parametric assumptions, since then the mediation formula prescribes only correct specification of the mediator's expectation.

In the light of these considerations, Vansteelandt *et al.* (2012b) recommended routine application of the imputation-based approach, especially when dealing with continuous mediators, since it avoids reliance on a model for the mediator. Despite this attraction, the imputation estimator does not come without limitations.

As in other imputation settings, one must pay due attention to coherent (or 'congenial') specification of the imputer's model and the analyst's model (i.e., in this case, the natural effect model) (Meng 1994). This might be particularly challenging for nonlinear outcome models. For instance, when using logistic regression to model binary outcomes, the imputation model may be difficult or impossible to match with the natural effect model (VanderWeele and Vansteelandt 2010; Tchetgen Tchetgen 2014). To limit the impact of potential model uncongeniality in terms of misspecification bias, Vansteelandt *et al.* (2012b) and Loeys *et al.* (2013) advocated the use of a sufficiently rich imputation model.[15] To this end, the **medflex** package allows users to fit an imputation model using generalized additive models or machine learning techniques, such as the ensemble learner as implemented in the **SuperLearner** package (Polley and van der Laan 2016).[16] Moreover, issues of uncongeniality can be avoided altogether by resorting to saturated natural effect models. In practice, models for conditional natural effects will rarely be saturated as either (some) baseline covariates or the exposure variable are continuous (or both). If the exposure is categorical, saturated models can be fitted for estimating population-average rather than stratum-specific natural effects (see Section 7). However, for observational data, as opposed to data from experiments where the exposure is randomly assigned, adjustment for confounding in population-average natural effect models requires inverse weighting for the exposure.[17]

[14] By default, the density function will correspond to the error distribution specified in the `family` argument for the mediator model (in turn specified via `neWeight`). QQ plots of the residuals can in this case be informative as to whether this parametric assumption is warranted for continuous mediators and can be obtained using the `qqnorm` function. The residuals can easily be obtained directly from the expanded dataset (as the working model is stored as an attribute in the expanded dataset object) by the command `residuals(expData)`.

[15] A 'minimal' imputation model should thus at least reflect the structure of the natural effect model (e.g., also including exposure-mediator interactions when these are postulated in the natural effect model as an interaction term between x and x^*) to avoid attenuation of the estimates of effects that were precluded from the imputation model.

[16] 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.

[17] Note that in both settings all baseline confounders still need to be adjusted for in the imputation model.

Second, as opposed to the weighting-based estimator, estimation by imputation requires modeling the mediator-outcome relation, which can be far from trivial whenever the exposure or baseline covariates are strongly associated with the mediator. In these scenarios, information about the effect of the mediator on the outcome may be sparse within certain strata defined by the exposure and covariates and, as a result, model misspecification may be difficult to diagnose and extrapolation bias becomes more likely (Vansteelandt 2012). Whenever increased concerns of model extrapolation arise, the weighting-based approach may be indicated, as extrapolation uncertainty will typically be more honestly reflected in the corresponding standard errors (Vansteelandt, Bekaert, and Claeskens 2012a).[18]

Finally, it can be argued that, for both estimation approaches, if the working model is correctly specified (either via generalized linear models or via more advanced techniques), a parsimonious (and possibly misspecified) natural effect model will still provide some summary result tailored to answer the practitioner's main research questions (Vansteelandt *et al.* 2012b; Loeys *et al.* 2013). Suppose, for instance, that the logistic regression models in equation (0) are correctly specified. Fitting a natural effect model of the form

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

to the expanded dataset using the imputation-based approach will then yield an estimated conditional natural indirect effect odds ratio of 1.143, which can be roughly considered as the mean conditional odds ratio across potential exposure levels (as depicted in Figure 1). If such an approach turns out to be unsatisfactory, users can again request residual plots to guide further model building and improve goodness-of-fit (by calling the `residualPlots` function). These diagnostics can be particularly helpful in the presence of certain non-linearities. For instance, when a continuous mediator is quadratic in the exposure, residual plots will indicate the need for a quadratic term for the indirect effect in the natural effect model, which will usually go unnoticed when fitting an imputation model for the outcome.

9.2. Missing data

As previously stated, when missingness occurs in the outcome, this is naturally dealt with when choosing the imputation-based approach, as missing outcomes in the original dataset are (by default) imputed in the expanded dataset, under the assumption that these outcomes are MAR (missing at random) given exposure, mediator(s) and baseline covariates.[19]

The weighting-based approach, on the other hand, is restricted to the analysis of complete cases and hence requires the more stringent MCAR (missing completely at random) assumption to hold in order to obtain unbiased estimation of the natural effect parameters. Whenever missingness occurs only in the outcome, we therefore advise to use the imputation-based approach. Alternatively, one might resort to multiple imputation, as also recommended if missingness occurs in either the exposure, mediator(s) or baseline covariates.

Moreover, although the use of population-average natural effect models can, in some settings, avoid issues concerning potential model uncongeniality, it is up to the researcher to decide whether stratum-specific or population-average effects are the target of the study.

[18] Extrapolation might also affect estimation in the natural effect model, primarily when baseline covariates and exposure are highly correlated. This concern holds for both the weighting- and imputation-based estimator, since both require regression adjustment for covariates to estimate conditional natural effects.

[19] It might be necessary to include additional covariates (that are both predictive of the outcome and missingness in the outcome, but are not included in the set of baseline covariates, C, that is chosen to meet assumptions A1–A4 to the imputation model to make the MAR assumption more plausible.

For instance, the `mice` function from the **mice** package (Van Buuren and Groothuis-Oudshoorn 2011) can be used to obtain multiply imputed datasets (stored in a `mids`-class object). The working model can in turn be fitted to each of these datasets by passing them (or rather the object containing these datasets) to the `with.mids` function, which also processes the function (i.e., either `neWeight` or `neImpute`) and expression that needs to be evaluated via the second argument. These steps are illustrated in the code below, in which `missdat` is a copy of the UPB dataset with artificially introduced missingness in each of the original variables.

```
R> library("mice")
R> library("mitools")
R> set.seed(123)
R> missdat <- UPBdata
R> for (i in 1:ncol(missdat)) {
+    missdat[sample(nrow(missdat))[1:10], i] <- NA
+  }
R> multImp <- mice(missdat, m = 10)
R> expData <- with(multImp, neWeight(negaff ~ factor(attbin) + gender +
+    educ + age))
```

Next, we use some functionalities from the **mitools** package (Lumley 2014) to fit a natural effect model (1) to each of the expanded multiply imputed datasets (stored in `expData$analyses`). The function `imputationList` can be used to transform the output containing these expanded datasets into a format that can be further passed to the `with.imputationList` function.

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

Finally, the results can be pooled by using the `MIcombine` function.

```
R> MIcombine(neMod1)
```

```
Multiple imputation results:
      with(expData, neModel(UPB ~ attbin0 + attbin1 + gender + educ +
    age, family = binomial("logit"), se = "robust"))
      MIcombine.default(neMod1)
                results         se
(Intercept) -0.84709412 0.73366115
attbin01     0.37963076 0.21874683
attbin11     0.34933687 0.08987945
gender2      0.34072605 0.24124907
educ2        0.14410803 0.49349023
educ3        0.39779433 0.50418124
age         -0.01097623 0.01299724
```

10. 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 allow for direct parameterization of the target causal estimands on their most natural scale. 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 any tools for assessing the sensitivity of one's results to possible violations of the identification assumptions of the causal estimands.

As mentioned in Section 8, 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 realm of natural effect models (such as a generic framework for conducting sensitivity analyses) will be added in updates of the package.

Acknowledgments

The authors would like to thank the Research Foundation Flanders (FWO) for financial support (Grant G.0111.12), Joris Meys for technical support during development of the **medflex** package and two anonymous reviewers whose feedback helped to improve this paper. Finally, we wish to thank Theis Lange, in particular, for helpful comments on an earlier draft of the paper and valuable suggestions during development of the package.

References

Albert JM (2008). "Mediation Analysis via Potential Outcomes Models." *Statistics in Medicine*, **27**, 1282–1304. doi:10.1002/sim.3016.

Albert JM (2012). "Distribution-Free Mediation Analysis for Nonlinear Models with Confounding." *Epidemiology*, **23**(6), 879–88. doi:10.1097/ede.0b013e31826c2bb9.

Albert JM, Nelson S (2011). "Generalized Causal Mediation Analysis." *Biometrics*, **67**(3), 1028–38. doi:10.1111/j.1541-0420.2010.01547.x.

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, San Francisco.

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. doi:10.1037/0022-3514.51.6.1173.

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. doi:10.1037/a0018933.

Canty A, Ripley BD (2016). ***boot**: Bootstrap R (S-PLUS) Functions*. R package version 1.3-18, URL https://CRAN.R-project.org/package=boot.

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." *The Stata Journal*, **11**(4), 479–517. doi:10.1002/9781119945710.ch17.

Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S (2015). "Causal Mediation Analysis with Multiple Causally-Ordered Mediators." *Biometrics*, **71**, 1–14. doi:10.1111/biom.12248.

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. doi:10.1007/s10896-012-9437-1.

Didelez V, Dawid AP, Geneletti S (2006). "Direct and Indirect Effects of Sequential Treatments." In R Dechter, T Richardson (eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pp. 138—-146. AUAI Press, Arlington, Virginia.

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*. 2nd edition. Sage, Thousand Oaks.

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 (2016). ***gam**: Generalized Additive Models*. R package version 1.14, URL https://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. doi:10.1111/bmsp.12028.

Hicks R, Tingley D (2011). "Causal Mediation Analysis." *The Stata Journal*, **11**(4), 1–15. doi:10.1007/978-1-4419-1764-5_8.

Holland PW (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association*, **81**(396), 945–960. doi:10.1080/01621459.1986.10478354.

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.

Hong G, Deutsch J, Hill HD (2015). "Ratio-of-Mediator-Probability Weighting for Causal Mediation Analysis in the Presence of Treatment-by-Mediator Interaction." *Journal of Educational and Behavioral Statistics*, **40**(3), 307–340. doi:10.3102/1076998615583902.

Hothorn T, Bretz F, Westfall P (2008). "Simultaneous Inference in General Parametric Models." *Biometrical Journal*, **50**(3), 346–363. doi:10.1002/bimj.200810425.

Huber M (2013). "Identifying Causal Mechanisms (Primarily) Based on Inverse Probability Weighting." *Journal of Applied Econometrics*, **29**(6), 920–943. doi:10.1002/jae.2341.

Iacobucci D (2012). "Mediation Analysis and Categorical Variables: The Final Frontier." *Journal of Consumer Psychology*, **22**(4), 582–594. doi:10.1016/j.jcps.2012.03.006.

IBM Corporation (2013). *IBM SPSS Statistics, Version 22.0*. IBM Corporation, Armonk. 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. doi:10.1037/a0020761.

Imai K, Keele L, Tingley D, Yamamoto T (2014). "Comment on Pearl: Practical Implications of Theoretical Results for Causal Mediation Analysis." *Psychological Methods*, **19**(4), 482–487. doi:10.1037/met0000021.

Imai K, Keele L, Yamamoto T (2010b). "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." *Statistical Science*, **25**(1), 51–71. doi:10.1214/10-sts321.

Imai K, Tingley D, Yamamoto T (2013). "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society A*, **176**(1), 5–51. doi:10.1111/j.1467-985x.2012.01032.x.

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. doi:10.1177/0193841x8100500502.

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. doi:10.1093/aje/kwr525.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**(1), 13–22. doi:10.2307/2336267.

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. doi:10.1080/00273171.2013.832132.

Lumley T (2014). ***mitools**: Tools for Multiple Imputation of Missing Data*. R package version 2.3, URL https://CRAN.R-project.org/package=mitools.

MacKinnon DP (2008). *Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, New York. doi:10.4324/9780203809556.

MacKinnon DP, Dwyer JH (1993). "Estimating Mediated Effects in Prevention Studies." *Evaluation Review*, **17**(2), 144–158. doi:10.1177/0193841x9301700202.

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.

Mayer A, Thoemmes FJ, Rose N, Steyer R, West SG (2014). "Theory and Analysis of Total, Direct, and Indirect Causal Effects." *Multivariate Behavioral Research*, **49**(5), 425–442.

Meng XL (1994). "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science*, **9**(4), 538–558. doi:10.1016/j.stamet.2006.03.002.

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. doi:10.1037/0022-3514.89.6.852.

Muthén B, Asparouhov T (2015). "Causal Effects in Mediation Modeling: An Introduction with Applications to Latent Variables." *Structural Equation Modeling*, **22**(1), 12–23. doi:10.1080/10705511.2014.935843.

Muthén LK, Muthén BO (2012). ***Mplus** User's Guide*. Muthén & Muthén, Los Angeles, 7th edition.

Pearl J (1995). "Causal Diagrams for Empirical Research." *Biometrika*, **82**(4), 669. doi:10.2307/2337329.

Pearl J (2001). "Direct and Indirect Effects." In J Breese, D Koller (eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01)*, UAI-01, pp. 411–420. Morgan Kaufmann Publishers, San Francisco.

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 & Sons, Chichester.

Pearl J (2014). "Interpretation and Identification of Causal Mediation." *Psychological Methods*, **19**(4), 459–481. doi:10.1037/a0036434.

Petersen ML, Sinisi SE, van der Laan MJ (2006). "Estimation of Direct Causal Effects." *Epidemiology*, **17**(3), 276–84. doi:10.1097/01.ede.0000208475.99429.2d.

Polley E, van der Laan M (2016). ***SuperLearner**: Super Learner Prediction*. R package version 2.0-21, URL https://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 (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robins JM (2003). "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." In P Green, NL Hjort, S Richardson (eds.), *Highly Structured Stochastic Systems*, pp. 70–81. Oxford University Press, New York.

Robins JM, Greenland S (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology*, **3**(2), 143–155. doi:10.1097/00001648-199203000-00013.

Robins JM, Richardson TS (2010). "Alternative Graphical Causal Models and the Identification of Direct Effects." In P Shrout (ed.), *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, pp. 103–158. Oxford University Press, Oxford, England.

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

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

Steen J, Loeys T, Moerkerke B, Vansteelandt S (2017). ***medflex**: Flexible Mediation Analysis Using Natural Effect Models*. R package version 0.6-1, URL https://CRAN.R-project.org/package=medflex.

Tchetgen Tchetgen EJ (2013). "Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis." *Statistics in Medicine*, **32**(26), 4567–80.

Tchetgen Tchetgen EJ (2014). "A Note on Formulae for Causal Mediation Analysis in an Odds Ratio Context." *Epidemiological Methods*, **2**(1), 21–31.

Tchetgen Tchetgen EJ, VanderWeele TJ (2014). "Identification of Natural Direct Effects When a Confounder of the Mediator Is Directly Affected by Exposure." *Epidemiology*, **25**(2), 282–91.

Tingley D, Yamamoto T, Hirose K, Keele L, Imai K (2014). "**mediation**: R Package for Causal Mediation Analysis." *Journal of Statistical Software*, **59**(5), 1–38. doi:10.18637/jss.v059.i05.

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. doi:10.1037/a0031034.

Van Buuren S, Groothuis-Oudshoorn K (2011). "**mice**: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1–67. doi:10.18637/jss.v045.i03.

van der Laan MJ, Petersen ML (2008). "Direct Effect Models." *The International Journal of Biostatistics*, **4**(1), 1–27. doi:10.2202/1557-4679.1064.

VanderWeele TJ (2011). "Causal Mediation Analysis with Survival Data." *Epidemiology*, **22**(4), 582–585. doi:10.1097/ede.0b013e31821db37e.

VanderWeele TJ (2013). "A Three-Way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects." *Epidemiology*, **24**(2), 224–232. doi:10.1097/ede.0b013e318281a64e.

VanderWeele TJ, Vansteelandt S (2009). "Conceptual Issues Concerning Mediation, Interventions and Composition." *Statistics and Its Interface*, **2**(4), 457–468. doi:10.4310/sii.2009.v2.n4.a7.

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. doi:10.1515/em-2012-0010.

VanderWeele TJ, Vansteelandt S, Robins JM (2014). "Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder." *Epidemiology*, **25**(2), 300–306. doi:10.1097/ede.0000000000000034.

Vansteelandt S (2012). "Understanding Counterfactual-Based Mediation Analysis Approaches and Their Differences." *Epidemiology*, **23**(6), 889–91. doi:10.1097/ede.0b013e31826d0f6f.

Vansteelandt S, Bekaert M, Claeskens G (2012a). "On Model Selection and Model Misspecification in Causal Inference." *Statistical Methods in Medical Research*, **21**(1), 7–30. doi:10.1177/0962280210387717.

Vansteelandt S, Bekaert M, Lange T (2012b). "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects." *Epidemiologic Methods*, **1**(1), Article 7. doi:10.1515/2161-962x.1014.

Vansteelandt S, VanderWeele TJ (2012). "Natural Direct and Indirect Effects on the Exposed: Effect Decomposition under Weaker Assumptions." *Biometrics*.

Yee TW, Wild CJ (1996). "Vector Generalized Additive Models." *Journal of the Royal Statistical Society B*, **58**(3), 481–493. doi:10.1007/978-1-4939-2818-7_19.

A. Link between estimators and the mediation formula

In this section we illustrate in more detail how natural effect models can be regarded as alternative formulations of 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.

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

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

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.

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

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.

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

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

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

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

Affiliation:

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

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

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
February 2017, Volume 76, Issue 11 *Submitted:* 2015-02-06
doi:10.18637/jss.v076.i11 *Accepted:* 2015-11-22