Health Serv Outcomes Res Method (2013) 13:86–108
DOI 10.1007/s10742-013-0106-5

Using mediation analysis to identify causal mechanisms in disease management interventions

Ariel Linden · Kristian Bernt Karlson

Received: 6 June 2012 / Revised: 26 February 2013 / Accepted: 13 March 2013 /
Published online: 24 March 2013

Abstract For over two decades, disease management (DM) has been touted as an intervention capable of producing large scale cost savings for health care purchasers. However, the preponderance of scientific evidence suggests that these programs do not save money. This finding is not surprising given that the theorized causal mechanism by which the intervention supposedly influences the outcome has not been systematically assessed. Mediation analysis is a statistical approach to identifying causal pathways by testing the relationships between the treatment, the outcome, and an intermediate variable that is posited to mediate the relationship between the treatment and outcome. This analysis can therefore help identify how to make DM interventions effective by determining the causal mechanisms between intervention components and the desired outcome. DM interventions can then be optimized by eliminating those activities that are ineffective or even counter-productive. In this article we seek to promote the application of mediation analysis to DM program evaluation by describing the two principal frameworks generally followed in causal mediation analysis; structural equation modeling and potential outcomes. After comparing several approaches within these frameworks using real and simulated data, we find that some methods perform better than others under the conditions imposed upon the models. We conclude that mediation analysis can assist DM programs in developing and testing the causal pathways that enable interventions to be effective in achieving desired outcomes.

A. Linden (✉)
Linden Consulting Group, LLC, 1301 North Bay Drive, Ann Arbor, MI 48103, USA
e-mail: alinden@lindenconsulting.org

A. Linden
Department of Health Management & Policy, School of Public Health,
University of Michigan, Ann Arbor, MI, USA

K. B. Karlson
SFI, The Danish National Centre for Social Research, Copenhagen, Denmark
e-mail: kbk@dpu.dk

K. B. Karlson
Department of Education, Aarhus University, Aarhus, Denmark

Keywords Disease management · Causal mediation analysis · Structural equation models · Potential outcomes · Observational studies

1 Introduction

For the past two decades, disease management (DM) has been promoted as an intervention capable of producing large scale cost savings for health care purchasers. Conceptually, cost savings are thought to be derived from reducing hospital admissions, emergency room visits, and the use of other costly health care services by helping patients adhere to self-care regimens, teaching them how to recognize acute exacerbations, and encouraging them to obtain recommended routine screening tests (Nelson 2012). In response, an array of commercial programs has been developed, resulting in the emergence of an industry estimated to be worth about $1 billion annually (Mattke et al. 2007).

However, these programs have struggled to deliver the anticipated cost savings. A recent Congressional Budget Office review of 34 Medicare DM demonstration projects stated that, "On average, the 34 programs had no effect on hospital admissions or regular Medicare expenditures (that is, expenditures before accounting for the programs' fees)," and, "After accounting for the fees that Medicare paid to the programs, […] Medicare spending was either unchanged or increased in nearly all of the programs" (Nelson 2012). Other evidence similarly points to the failure of commercial DM programs to achieve medical cost savings. A broader Congressional Budget Office review of the DM literature concluded that "there is insufficient evidence to conclude that DM programs can generally reduce the overall cost of health care services" (Congressional Budget Office 2004). This report was followed by several additional systematic reviews that arrived at similar conclusions (Ofman et al. 2004; Goetzel et al. 2005; Mattke et al. 2007).

Despite this evidence, payors in the private sector continue to purchase DM services and DM continues to be discussed as a viable approach to achieving cost savings (Matheson et al. 2006; Mays et al. 2007). It is therefore critically important to determine why the DM model has failed to result in cost savings to-date and how it can be modified going forward. A logical place to begin is by examining the various components of the DM intervention and the associated causal mechanisms that could lead to a reduction in medical costs. For example, a core component of the standard DM intervention is for nurses to talk to patients by phone to increase their self-efficacy such that patients feel empowered to make behavioral changes that will improve their health. However, nurses employed by DM programs are rarely pre-screened to assess aptitude and/or acceptance of a patient-centric model based on behavior change science (Butterworth and Andersen 2011; Miller and Rose 2009), nor are they adequately trained in behavior change approaches (Linden and Roberts 2004; Linden et al. 2006), nor is there widespread use of validated and standardized tools to assess the nurses' proficiency and ensure fidelity to those evidence-based behavioral change approaches (Butterworth and Andersen 2011). Without a structured training and continual assessment process, it is unlikely that these nurses will achieve the proficiency level necessary to improve patients' self-efficacy to self-manage their chronic illness (Marks et al. 2005; Miller and Rose 2009). Patients with poor self-efficacy will not likely change their health behaviors or interact with their providers more effectively (Bodenheimer et al. 2002; Marks et al. 2005). As a result, their health care utilization and costs should not be expected to change.[1]

[1] Linden and Adler-Milstein (2008) highlight many additional factors that explain why the standard DM approach is not effective, based on the best available evidence from the literature.

This example sets up two important and interrelated points about the design and assessment of DM programs. First, at the outset, the causal mechanism(s) by which the intervention is hypothesized to influence the outcome should be specified—leveraging both content expertise and empirical evidence. This approach should make it clear that it is *possible* to achieve the intended aims. A specified casual pathway has a second and arguably more important benefit of enabling an evaluation that assesses whether the intervention is in fact working through the hypothesized pathways. This involves mediation analysis, which entails identifying intermediate variables which lie on the casual pathway between treatment and outcome and then assessing whether the treatment impacts the mediator variable which in turn effects change in the outcome. If a program is effective, mediation analysis confirms that it is operating in the anticipated way. However, even in ineffective programs, mediation analysis is far more useful than the standard analytic approach, which ignores the casual pathway, because mediation analysis may identify where the casual pathway is breaking down. Such information directly informs how best to target program improvement efforts.

While mediation analysis is increasingly being used to study causal mechanisms across a variety of disciplines and settings, the concept has yet to be adopted by the DM industry. Broader use of this technique could be immensely beneficial as it would elucidate where the DM model is failing. Therefore, in this paper, we introduce readers to the concept of mediation analysis, with an emphasis on DM interventions. To that end, we describe several analytic methods generally used to conduct mediation analysis, including both traditional structural equation modeling (SEM) methods popularized by Baron and Kenny (1986) and recently introduced approaches based on the potential outcomes framework originally proposed by Rubin (1974, 1978), together with key assumptions required to interpret mediation results as causal. We then use both actual and simulated data to compare the results generated from the various mediation analysis methods introduced. Finally, we discuss the implications of our findings and provide direction for researchers wishing to conduct mediation analysis as part of a more comprehensive and informative evaluation of DM program effectiveness.

2 The SEM approach to mediation analysis

The basic conceptual framework of a mediation process with a single mediator is illustrated in Fig. 1. As shown, Treatment (T) can impact the outcome (Y) either indirectly via the mediator (M) or directly. In health management interventions we may expect a significant proportion of the effect to be direct, since there are likely to be myriad variables not observed through the mediated pathway (including other unmeasured mediators). Thus, the total treatment effect is the sum of both direct and indirect effects. These associations can be expressed statistically using the following set of linear regressions:

$$\mathrm{Y}_i = \alpha_1 + c\mathrm{T}_i + \beta_1\mathrm{X}_i + \varepsilon_i \tag{1}$$

$$\mathrm{M}_i = \alpha_2 + a\mathrm{T}_i + \beta_2\mathrm{X}_i + \varepsilon_i \tag{2}$$

$$\mathrm{Y}_i = \alpha_3 + b\mathrm{M}_i + c^{'}\mathrm{T}_i + \beta_3\mathrm{X}_i + \varepsilon_i \tag{3}$$

Equation (1) is a standard outcomes model estimating the average total effect of the intervention by regressing the outcome Y on the treatment variable T and one or more pre-intervention characteristics X. Equation (2) represents the *a* pathway in Fig. 1 in which the

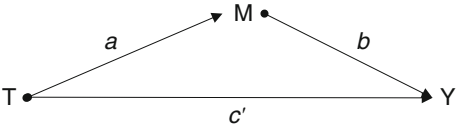

Fig. 1 The conceptual mediation model with a single mediator. *T* treatment assignment, *M* mediator, *Y* outcome. SEM coefficients are represented by *a*,*b*,*c*′

mediator M is regressed on T and X. Equation (3) provides both the *b* and *c*′ pathways indicated in Fig. 1 by regressing the outcome on T, M, and X. Thus, *b* represents the effect of M on Y controlling for T and X, and *c*′ represents the effect of T on Y, controlling for M and X.

With Equations (1–3), mediation effects can be estimated using one of two methods. The "product of coefficients" method refers to the product of the *a* and *b* pathways (*ab*) while the "difference in coefficients" method subtracts the direct effect *c*′ from the total effect *c* to derive at the indirect effect ($c - c'$) (Stolzenberg 1980; MacKinnon et al. 2002). In the absence of Eq. (1), the total effect can be computed as the sum of the direct and indirect effects ($c = ab + c'$). Finally, in mediation analysis, researchers are often not only interested in point estimates, but also in the extent to which a variable mediates a relationship. A natural quantification is the mediation percentage, which is calculated as the ratio of the indirect effect to the total effect.

Given that the estimation of indirect effects utilizes values generated from two separate regression models, correct specification of standard errors and tests of significance can be performed using specialized procedures (Freedman and Schatzkin 1992; Sobel 1982) or bootstrapping (Efron and Tibshirani 1993; Shrout and Bolger 2002).

The SEM approach for studying the effect of multiple mediators, popularized by Baron and Kenny in 1986, is a straightforward extension of the single mediator model (Alwin and Hauser 1975; MacKinnon 2008, Chap. 5; Bollen 1989; Duncan 1966). Basically, each mediator is regressed separately on the treatment variable and pre-intervention characteristic (T and X, respectively), and then the outcome model regresses the outcome on all the mediators as well as on T and X.

2.1 Assumptions required for interpreting mediation effects as causal in the SEM approach

While intuition may suggest that mediation effects can be readily interpreted when studied within a randomized controlled trial (RCT) context, in fact only the path between treatment and mediator can be considered causal under this design. Because individuals are not randomly assigned to the various mediator levels, any differences found in the outcome may be a result of self-selection at the level of the mediator or confounding that is introduced post-treatment (Holland 1988; Jo 2008; Sobel 2008).

Given the potential for bias that may arise at either treatment assignment (in observational studies) and/or mediator stage (both in RCTs and observational studies), the primary assumption required for causal interpretation of mediational processes is that of *sequential ignorability* (Imai et al. 2010a, b, c). The first part of the sequence assumes that treatment is independent (ignorable) of potential mediators and outcomes, allowing for a causal interpretation of path *a*. In an RCT, this assumption is ensured by randomization, while in observational studies the assumption can be met when conditioning on observed pre-

intervention covariates [noted as X in Eqs. (1–3)] leads to no residual confounding (Rosenbaum and Rubin 1983).[2] The second part of the sequence assumes that the level of the mediator is independent of the potential outcomes (in both RCTs and observational studies) conditional on treatment and observed pre-intervention characteristics (Imai 2010a, b, c). In other words, after conditioning on observable pre-intervention characteristics X and the actual treatment assignment T, we now assume that the mediator status is *as good as randomized*. This assumption allows us to interpret path *b* as causal because individuals within each treatment group attaining different levels of the mediator should be similar and thus can be compared.[3] Similarly, path *c* can be causally interpreted because individuals across different treatment groups attaining the same level of the mediator should also be similar and comparable (Jo 2008). It is important to note that the sequential ignorability assumption cannot be directly tested from the data, and therefore to a large extent, the researcher must present convincing arguments that the assumption holds when estimating causal mediation effects.

Another assumption often maintained in causal mediation analysis using SEM is that of no interaction between treatment and mediator. This assumption states that treatment has a constant direct effect (path c') on the outcome regardless of mediator level, and that the effect of the mediator on the outcome (path *b*) is constant across different treatment assignments (Jo 2008). Whenever the no-interaction assumption does not hold, a standard result in the literature using SEM is that the indirect effect depends on the level of the treatment variable (Stolzenberg 1980; Kraemer et al. 2008), i.e., the (average) indirect effect differs among treated and untreated (Judd and Kenny 1981; Imai et al. 2010a, b, c). Kraemer et al. (2008) suggest that the constant effect assumption is rather unrealistic and recommend including an interaction term (treatment X mediator) in the outcome model to eliminate the requirement for this assumption:

$$\mathrm{Y}_i = \alpha_3 + b\mathrm{M}_i + c^{'}\mathrm{T}_i + d\mathrm{TM}_i + \beta_3\mathrm{X}_i + \varepsilon_i \tag{4}$$

In this alternative specification, a statistically significant interaction term *d* indicates that the relation between M and Y differs by treatment group, thus violating the constant effect assumption (and possibly producing different direct and indirect effect estimates). Conversely, one might assume that if the interaction term is not-statistically significant then the constant effect assumption holds. However, Glynn (2012) illustrates that this assumption is somewhat misleading, because the mere inclusion of the interaction term (even when the coefficient is not statistically significant), may generate substantially different direct and indirect estimates than those produced by the standard model without the interaction term. Referring back to our DM example in which patient self-efficacy could be examined as a mediator of the relationship between DM intervention and medical costs (Lorig and Holman 2003), it is likely that the treatment and control groups would differ on their mean levels of self-efficacy, either as a result of the effectiveness of the intervention or possibly due to self-selection bias and/or confounding. In turn, this would likely impact the outcome differentially between groups. A statistically significant interaction term should be explored further to determine the source of the significance. MacKinnon (2008, p. 280) recommends using contrasts to test the significance of the *b* coefficient relating M to Y at the different levels of T as well as reviewing visual displays of the data. As a partial

[2] In observational studies the assumption of no residual confounding or no selection bias cannot be tested, and so causal effects are only identified under this assumption.

[3] "Similar" refers to comparability on both observed and unobserved characteristics, because we assume no residual confounding (ignorability) once we have conditioned on pre-treatment covariates and treatment.

solution to an interaction effect, Glynn (2012) suggests restricting inferences to sub-populations of interest. This approach is expanded by Imai et al. (2010a, b, c) in which the mediation effect can be estimated separately for the treatment and control group and then pooled as a weighted average estimate. Perhaps most importantly it should be noted that a statistically significant treatment X mediator interaction will not negate findings of mediation but may in fact provide richer more detailed information about an observed mediation effect (Kraemer et al. 2008; MacKinnon 2008, p. 295).

A final assumption required for causal interpretation of mediational processes using the SEM approach is that the relationship between the continuous mediator and outcome is linear (Sobel 2008; Jo 2008). In other words, we assume that that the outcome value linearly increases (or decreases) as the mediator value increases (or decreases) (Jo 2008). As will be discussed in the next section, this becomes even more problematic when the mediator and/or outcome variable is categorical.

2.2 Categorical outcome or mediator variables

The standard SEM approach (Eqs. 1–3) utilizes ordinary least squares regression with the understanding that the mediator and outcome variables are continuous. However, in DM evaluations some outcome or mediator variables are binary (yes/no), such as the receipt of an appropriate lab test, quitting smoking, or filling a prescription, which often are modeled using logit or probit regression.[4]

Contrary to the linear models described thus far where the mediation effect is estimated by $c - c'$, in nonlinear probability models for categorical outcomes (i.e., logit or probit), $c - c'$ does not recover the mediation effect because of a rescaling or attenuation of the model in Eq. 3 that occurs when the mediator variable has an independent effect on the outcome (Winship and Mare 1983, 1984; Wooldridge 2002; Cramer 2003). In other words, in these models the inclusion of the mediator variable M in Eq. 3 will alter the coefficient c' merely if M is correlated with Y, thereby conflating mediation with rescaling, which results in biased mediation effects using $c - c'$.

Several approaches have been proposed to resolve this issue. MacKinnon and Dwyer (1993) suggested two approaches using the method of Y-standardization, which rescales coefficients to be measured in standard deviations of the latent outcome variable assumed to underlie the binary outcome variable, giving coefficients an interpretation similar to that found for standardized coefficients in linear models (McKelvey and Zavoina 1975; Winship and Mare 1983, 1984; Long 1997). The first approach standardizes the coefficients linking T to M and M to Y (a and b) and then applies the "product of coefficients" method to these coefficients (ab). In situations where M is continuous, the T–M relationship is obtained from the standardized coefficient of a linear model, whereas in situations where M is a categorical variable, the coefficient is obtained by standardizing the coefficient of a logit or probit model. The second approach follows Winship and Mare (1983, 1984) and standardizes the coefficients of the treatment variable in the model with and without the mediator (i.e., c and c') and then applies the "difference-in-coefficients" method to these standardized coefficients. This approach thereby does not model the T–M relationship directly, but rather compares coefficients in the model for the outcome measured on the same scale.

[4] Mediators or outcomes may also be ordered, such as rating of perceived health status or satisfaction on a Likert-type scale (e.g., 1 through 5), which can be modeled using ordered logit or probit models (which are natural extensions of the logit or probit models).

More recently an alternative approach to overcoming the limitations of evaluating mediation in logit and probit models has been introduced (Karlson et al. 2012; Karlson and Holm 2011).[5] The method by Karlson et al. (2012) allows comparisons of total and direct effects of treatment variables that are unaffected by rescaling or attenuation bias. It exploits the properties of a rescaled logit regression to generate an estimate of $c - c'$ which is unaffected by rescaling bias. Further work clarified that this estimate is equal to the product between, (a) the effect of the mediator on the binary outcome in a logit or probit model (controlling for treatment), and (b) the effect of treatment on the mediator in a linear regression model. In their method, the T–M relationship is always modeled using a linear model, meaning that for categorical M, linear probability models are used. The results of Karlson et al. (2012) suggest, first, that the much-discussed difference in results produced by the "difference in coefficients" method and the "product of coefficients" method in logit and probit models disappear once the latter method is applied using the rescaled logit regression and, second, an equivalence between the decomposition principles of linear models and nonlinear probability models. These conclusions are consistent with the results reported on the "product of coefficients" method in nonlinear probability models by MacKinnon et al. (2007). As in the linear SEM case, the method by Karlson et al. (2012) also allows multiple mediators and the inclusion of pre-treatment covariates. Whenever the mediator is binary, all methods discussed thus far use a linear probability model for modeling the T–M relationship, except for the first Y-standardization approach, which uses a standardized logit or probit coefficient.

3 Potential outcome approaches to mediation analysis

Given the challenges of satisfying the assumptions of the SEM approach, in particular ignorability in the relationship between M and Y, a parallel stream of research has approached mediation analysis using the potential outcomes framework (Rubin 1974, 1978) to clarify the assumptions under which the SEM approach allows for causal interpretation (Holland 1986, 1988; Robins and Greenland 1992; Pearl 2001; Jo 2008; Sobel 2008; Imai et al. 2010a, b, c). While the SEM approach and the potential outcomes framework are similar in many respects, the latter is a more general framework. It allows for a broader, "nonmodel-based" understanding of causal effects while allowing the SEM approach to be considered a special case of this broader framework (Pearl 2012).

To illustrate this framework, assume a DM program where $Y_i(1)$ represents the outcome of an individual who was assigned to the intervention and $Y_i(0)$ represents the outcome if that individual was assigned to the control group. The individual level treatment effect is $Y_i(1) - Y_i(0)$, or the difference in outcomes experienced by the individual after being exposed to both treatment and control conditions. For any individual only one of these outcomes is observed, and so researchers generally estimate average treatment effects at the group level, relying on an equivalent control group to represent the counterfactual outcome. This strategy is the first part of the sequential ignorability assumption described earlier which expects treatment assignment to be independent of potential mediators and outcomes.[6]

[5] Breen et al. (Forthcoming) develop the method further and suggest some further identities we refer to here.

[6] An additional assumption required here for causal inference is that each individual's potential outcomes are unrelated to the treatment status of any other individual under study (Rubin 1978; Manski, Forthcoming).

To illustrate how the potential outcomes framework is extended under mediation analysis, we broaden the notation from above as follows: assume a $M_i(1)$ represents the mediator value of an individual who was assigned to the intervention and $M_i(0)$ represents the mediator value if that individual was assigned to the control group. We can combine the outcome and mediator variables such that $Y_i(1, M_i(1))$ describes the outcome of an individual assigned to the intervention group who achieves a mediator level that would be realized under that treatment condition and $Y_i(0, M_i(0))$ describes the outcome of this individual if assigned to the control group who achieved a mediator level realized under the control condition (Imai et al. 2010a, b, c).

Following definitions provided by Pearl (2001) and Robins and Greenland (1992) (with notation from Imai et al. 2010a, b, c) the direct effect is represented as:

$$\text{Direct effect } = Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \tag{4}$$

which can be interpreted as the effect of treatment on the outcome holding the mediator at the level of the control condition (or stated differently - the effect of treatment on the outcome not conveyed via the mediator). This is equivalent to coefficient *c'* in the SEM approach (Eq. 3). Similarly, the direct effect of treatment on the outcome holding the mediator at the level of the treatment group can be estimated by setting both treatment indicators for the mediator to $M_i(1)$. The indirect effect is defined as:

$$\text{Indirect effect} = Y_i(1, M_i(1)) - Y_i(1, M_i(0)) \tag{5}$$

which represents the effect of changing the mediator level (from that observed in the control group to that observed in the treatment group) on the outcome, holding the treatment assignment constant (set here to the treatment group). By setting the treatment status constant, we isolate the effect of the mediator on the outcome while controlling for other possible effects induced by the treatment. Similarly, we can estimate the indirect effect on the outcome holding the treatment assignment at the level of the control group by setting both treatment indicators for the outcome to $Y_i(0)$. This is equivalent to the product of coefficients (*ab*) in the SEM approach with no interaction. The total effect of the DM intervention on the outcome is:

$$\text{Total effect} = Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \tag{6}$$

which is equivalent to summing the direct and indirect effects[7]:

$$\text{Total effect } = (Y_i(1, M_i(0)) - Y_i(0, M_i(0))) + (Y_i(1, M_i(1)) - Y_i(1, M_i(0))) \tag{7}$$

As before, for each individual we observe only one outcome, and in addition, we only observe one mediator value. Thus, in estimating mediation effects using the potential outcome approach we rely on an equivalent control group to represent the potential values of the mediator and outcome. It is quite possible that direct and indirect effects (per Eqs. 4 and 5) will differ when the group assignment is set to equal the treatment versus that of control. As Glynn (2012) notes, when the indirect effect is estimated setting the assignment variable equal to the treatment group, the resulting estimate (i.e., the average indirect effect on the treated) is analogous to the sample average treatment effect on the treated (SATT).

Footnote 6 continued
In an evaluation of a DM intervention, this assumption could be violated if members of the same household were enrolled in the intervention, possibly influencing the outcomes of one another.

[7] This holds under the no interaction assumption in which direct and indirect effects are assumed to be identical between treatment and control groups (see Imai et al. 2010, p. 312).

By extension, when the indirect effect is estimated setting the assignment variable equal to the control group, the resulting estimate (i.e., the average indirect effect on the controls) is analogous to the sample average treatment effect on the controls (SATC). Program evaluators conducting mediation analysis may be more interested in one estimator over the other.[8] However, the approach by Imai et al. 2010a, b, c, which we describe below, allows for the mediation effect to be estimated separately for the treatment and control group and then pooled as an average estimate across the two conditions. Moreover, in linear models and assuming no interaction between T and M, mediation effects on treated and untreated are identical.

Several methods have been proposed to estimate mediation effects motivated by the potential outcomes framework. These include semi- and non-parametric estimation procedures (Imai et al. 2010a, b, c; Pearl 2001, 2011; Hafeman and Schwartz 2009), matching on the propensity score (Hill et al. 2003), weighting approaches (Peterson et al. 2006; VanderWeele 2009; Hong 2010), principal stratification (Frangakis and Rubin 2002; Jo 2008; Jo et al. 2011) and the G-computation algorithm (Robins and Greenland 1992). Many of these approaches overlap conceptually, or serve as natural extensions of other approaches. For example, the propensity score (Rosenbaum and Rubin 1983) features prominently as a way of addressing the sequential ignorability assumption, in either a stand-alone procedure (Hill et al. 2003) or as a basis for weighting and principal stratification (Peterson et al. 2006; VanderWeele 2009; Hong 2010; Jo et al. 2011).

We briefly describe two methods chosen from the vast array of those available that are well-suited to DM evaluations. The first approach described by Imai et al. 2010a, b, c, is by far the most flexible, designed to accommodate most mediator/outcome variable types that a researcher will likely come across in practice.[9] In the second approach, VanderWeele (2009) extends the propensity score-based weighting technique used for causal inference popularized by Robins (1998) and Robins et al. (2000). This approach is likely to appeal to researchers already accustomed to using propensity score-based weighting approaches in program evaluation.[10]

Imai et al. (2010a, b, c) propose both parametric and nonparametric procedures for estimating average mediation effects. Here we describe the parametric procedure for estimating the mediation effect in the treatment group (as notated in Eq. 5) (see also Hicks and Tingley 2011). First, the mediator is regressed on the treatment variable and other pre-intervention covariates as in Eq. 2. Second, two individual-level predictions from this model are stored, once setting the treatment status to equal treatment, and then again as control. In other words, this step predicts both the actual and counterfactual level of the mediator for each individual in the treatment group. Third, the outcome is regressed on the mediator and other pre-intervention covariates for the treatment arm only. Fourth, using the regression formula estimated in the prior step, two individual-level potential outcome predictions are stored, once when replacing the actual mediator value with that of the predicted mediator value for the treated condition, and then again using the predicted counterfactual mediator value for the treatment group (both values were generated in Step 2). The average indirect effect (on the treated group) is estimated in the fifth and final step by simply calculating the average difference between the outcome predictions using the actual and counterfactual values of the mediator. This iterative procedure is exemplified by

[8] See Morgan and Todd (2008) for a discussion on assessing the effects of the ATT versus ATC estimators.

[9] The method is implemented in both the R Language (Imai et al. 2010) and in Stata (Hicks and Tingley 2011).

[10] See Linden and Adams (2010a, b) for a description of these techniques used in the DM context.

Eq. 5, and can be easily replicated for the control group, accordingly. Finally, the percentage of the total effect that is mediated can be calculated by dividing the result of Eq. 5 by Eq. 6 (or 7).[11]

VanderWeele (2009) proposes a different method which involves generating separate propensity-score based weights (Robins 1998; Robins et al. 2000) for the mediation and outcome models, and then estimating weighted regressions within the usual SEM framework. What follows is a description of the process for estimating the indirect mediation effect when both treatment and mediator variables are binary (we later discuss how other variable types are handled).

The first propensity score is estimated by regressing the treatment variable on the observed pre-intervention covariates using a logit or probit model. This propensity score is defined as the probability of assignment to the treatment group conditional on covariates (Rosenbaum and Rubin 1983), and controls for pre-intervention differences between treatment participants and non-participants. Second, the inverse probability of treatment weight (IPTW) is computed by giving treatment group participants a weight equal to the inverse of the estimated propensity score (1/propensity score), and non-participants a weight equal to the inverse of 1 minus the estimated propensity score (1/(1-propensity score)) (Robins 1998; Robins et al. 2000). This weight is used directly in a weighted regression for the mediator model (Eq. 2).

In the third step, a second propensity score is estimated by regressing the mediator on the treatment variable and observed intervention covariates using a logit or probit model. This propensity score can be defined as the probability of obtaining a given mediator level conditional on treatment group assignment and pre-intervention covariates. Fourth, the associated IPTW is computed by giving individuals obtaining a mediator value of 1 a weight equal to the inverse of the estimated propensity score (1/(1-propensity score)), and individuals obtaining a mediator value of 0 a weight equal to the inverse of 1 minus the estimated propensity score (1/1-propensity score). Fifth, the two weights are multiplied together and then used directly in a weighted regression for the outcome model (Eq. 3), and can be considered a different method of adjusting for pre-treatment confounders compared to the standard regression framework. In essence, this composite weight is a means of addressing both parts of the sequential ignorability assumption. One point worthy of note is that when covariate adjustment is not required or not possible, weights are not computed and thus this approach defaults back to the standard SEM approach.

As described generally above, the propensity score is estimated by regressing either the treatment variable or mediation variable on a set of covariates. More specifically, however, the choice of regression model depends on the variable type of the treatment or mediation variable used in the equation. For example, either logit or probit models are typically used to estimate the propensity score when the treatment and/or mediation variable is binary. When treatment and/or mediation variables are ordered, ordinal logit/probit models can be used to estimate the propensity score (Joffe and Rosenbaum 1999; Lu et al. 2001; Zanutto et al. 2005), and in the case of a continuous treatment and/or mediation variable, application of the generalized propensity score (Hirano and Imbens 2004; Imai and van Dyke 2004) may be considered.

[11] However, in the software (Hicks and Tingley 2011), the reported "percent mediated" is the median of a simulated distribution of "percent mediated," and thus may not provide the same result as that derived by dividing Eq. 10 by Eq. 11. We return to this point in our Monte Carlo study.

4 Comparison of approaches

As the number of methods for mediation analysis has increased in recent years, the question arises of how the methods compare. To help address this question, we apply a range of methods to the JOBS II dataset used in Imai et al. (2010a, b, c) and to simulated data in a Monte Carlo study. In both applications we compare the linear SEM approach suggested by Baron and Kenny (1986), the approach suggested by Imai et al. (2010a, b, c), the approach suggested by Karlson et al. (2012) (KHB), the two approaches based on Y-standardization suggested by MacKinnon and Dwyer (1993), and the approach using inverse probability of treatment weighting (IPTW) suggested by VanderWeele (2009).

4.1 JOBS II data

In our first comparison of approaches we use the JOBS II study which Imai et al. (2010a, b, c) used for illustrating their mediation approach. We briefly reiterate their description of the study (p. 310). JOBS II is a randomized experiment in which unemployed workers were allocated to either a treatment or control group. The treated participated in a job skills workshop, whereas the controls received a booklet giving job search tips. The outcome of interest was a post-treatment continuous measure of depression, while the mediator of interest was a continuous measure of job search self-efficacy. The mediator thus represents the mechanism through which the treatment effect is hypothesized to be delivered: Workshop participation strengthens self-efficacy which in turn reduces depression. The study also provided a range of pre-treatment control variables: age, sex, marital status, previous occupation, income, education, a measure of economic hardship, and a pre-treatment measure of depression.

We use JOBS II for two purposes. First, we report total, direct, and indirect (mediation) effects estimated with the mediation approaches. Because both outcome and mediator are continuous measures in JOBS II, we can construct binary versions of the continuous measures to illustrate the application of the methods on other outcome types. In the first analysis we study two situations often met in applied research, in which the outcome is either continuous or binary. In both situations, we use a continuous mediator. Second, we report mediation percentages, i.e., the indirect effect over the total effect, in the four situations that would be typically met in applied research.[12] If Y denotes the dependent variable and M the mediator variable, then the four situations are: Y-continuous M-continuous, Y-continuous M-binary, Y-binary M-continuous, Y-binary M-binary. In both analyses the binary versions are constructed by grouping respondents according to whether or not they pass a certain threshold on the depression variable (Y) and on the self-efficacy variable (M), respectively.[13] As in Imai et al. (2010a, b, c), we adjust all models for the pre-treatment control variables.

[12] MacKinnon et al. (1995) demonstrate that mediation percentages are unstable for smaller sample sizes. We nevertheless choose to report these percentages here, because they are widely used in applied research and because they provide a sensible metric for comparing results in nonlinear probability models in which point estimates of total, direct, and effects are identified up to an arbitrary scale.

[13] We dichotomize these variables strictly to illustrate the modeling approach. In practice, converting continuous variables to dichotomous or categorical variables should be avoided, as it leads to a loss of information and reduces power (Royston et al. 2006).

Table 1 Comparison of mediation two approaches using JOBS II data from Imai et al. (2010a, b, c) with a continuous outcome and a continuous mediator

	Total effect			Direct effect			Indirect effect		
	Est.	95 L-CL	95 U-CL	Est.	95 L-CL	95 U-CL	Est.	95 L-CL	95 U-CL
Linear SEM/KHB Linear	−0.060	−0.143	0.023	−0.042	−0.126	0.041	−0.018	−0.039	0.003
Imai et al.: Linear-Linear	−0.060	−0.119	0.003	−0.042	−0.121	0.040	−0.018	−0.041	0.002

The linear SEM which is equivalent to the method of Karlson et al. (2012) applied to a continuous outcome and the method by Imai et al. (2010a, b, c). Values represent coefficients for estimated total, direct, and indirect effects

See Table 3 for method descriptions

4.1.1 Total, direct, and indirect effects

Table 1 reports the total, direct, and indirect effects in the scenario with a continuous outcome and a continuous mediator where we assume linearity and no interaction between treatment and mediator.[14] The first column uses the method by Baron and Kenny (1986), while the second column contains the estimates using the method by Imai et al. (2010a, b, c). The point estimates of the total, direct, and indirect effects are identical between the two methods (up to three decimals), which is what we would have expected given the results in Imai et al. (2010a, b, c). The confidence intervals suggest that the method by Imai et al. (2010a, b, c) is slightly more efficient than the Baron and Kenny approach (1986).

In Table 2 we report effects when the outcome is binary and the mediator is continuous. In the first row we report the results using a linear probability model for the outcome model, meaning that the effects are measured on the probability margin. For example, the estimate of the total treatment effect is −6.8 percentage points. The estimates produced by the method by Imai et al. (2010a, b, c), using a linear probability model for outcome, reported in the third row, are identical to those obtained by those in the first row (up to three decimals). In the fourth row in Panel B we report the results using the probit for the outcome in the Imai et al. (2010a, b, c) approach. Imai et al. (2010a, b, c) define these effects on the probability margin and, given the nonlinearity of the probit link, mediation effects can differ between treated and untreated, even when the effects in the underlying latent model do not. We therefore report the effects for the treated (T = 1) and untreated (T = 0), although results in Panel B are near-identical for these two groups. Compared to the previous approaches, the effects are smaller, although not much. For example, the direct effect for the treated or untreated is −4.7 percentage points, compared to −5.6 percentage points with the Baron and Kenny (1986) approach. In contrast to the other approaches, the indirect effect for the treated is 0.1 percentage points larger than for the untreated, suggesting the sensitivity of the method by Imai et al. (2010a, b, c) to the nonlinearity of the probit link. However, the magnitude of the indirect effect is virtually the same across all approaches (−0.012).

In the seventh and eighth row in Table 2 we turn to the method by Karlson et al. (2012). This method is derived using the latent linear model assumed to underlie the probit (or

[14] We first estimated an interaction model which produced a non-significant interaction effect ($p = 0.230$, CI: −0.044, 0.18), followed by a review of the contrasts between groups at each level of the mediator which supported the no-interaction effect assumption.

Table 2 Comparison of mediation approaches using JOBS II data from Imai et al. (2010a, b, c), with a binary outcome and continuous mediator

	Total effect			Direct effect			Indirect effect		
	Est.	95 L-CL	95 U-CL	Est.	95 L-CL	95 U-CL	Est.	95 L-CL	95 U-CL
Linear SEM	−0.068	−0.128	−0.008	−0.056	−0.116	0.004	−0.012	−0.027	0.002
Imai et al.:									
Linear–Linear	−0.068	−0.112	−0.022	−0.056	−0.113	0.004	−0.012	−0.028	0.001
Probit-Linear T = 1	−0.059	−0.106	−0.013	−0.047	−0.106	0.008	−0.012	−0.029	0.001
Probit-Linear T = 0	−0.059	−0.106	−0.013	−0.047	−0.106	0.007	−0.011	−0.029	0.001
KHB									
Probit	−0.207	−0.405	−0.010	−0.166	−0.364	0.032	−0.041	−0.090	0.008
Probit APE	−0.063	−0.122	−0.003	−0.050	−0.110	0.009	−0.012	−0.027	0.003
Y-standardization A: Probit	−0.092	−0.183	0.0002	−0.074	−0.163	0.016	−0.018	−0.040	0.004
Y-standardization B: Probit	−0.188	−0.363	−0.014	−0.142	−0.312	0.027	−0.046	−0.090	−0.002

Values represent coefficients for estimated total, direct, and indirect effects

Linear SEM uses linear models for both outcome and mediator model, as in Baron and Kenny (1986). Imai et al. refers to the method by Imai et al. (2010a, b, c), calculated with the user-written Stata© command *medeff* (Hicks and Tingley 2011); the combination of link functions refer to the outcome and mediator model, respectively. KHB probit uses the user-written Stata© command *khb* (Kohler et al. 2011), which implements the method by Karlson et al. (2012); it uses the probit link function. Y-standardization A uses the method suggested by Winship and Mare (1984) and applied in McKinnon and Dwyer (1993) as a "difference in coefficients" method, using the probit link function for the outcome model. Y-standardization B uses the "product of coefficient" in McKinnon and Dwyer (1993, pp. 151), which uses Y-standardization for binary mediators, and which is implemented in the user-written Stata© command *binary_mediation*; in the case with continuous mediators, this method defaults to the KHB. IPTW uses the method described by VanderWeele (2009); computation of inverse probability of treatment weights is based on the logit model; in the Y-binary M-binary scenario the probit model is used for the outcome equation. Confidence intervals for Y-Standardization A and B and for the indirect effect of KHB Probit APE are calculated using the bootstrap (1,000 replications)

Table 3 Comparison of mediation approaches using JOBS II data from Imai et al. (2010a, b, c)

Method	Y-continuous M-continuous	Y-continuous M-binary	Y-binary M-continuous	Y-binary M-binary
Linear SEM	27.183	38.626	17.734	25.042
Imai et al.				
Linear-Linear	22.352	32.861	17.770	24.958
Linear-Probit	–	27.484	–	–
Probit-Probit	–	–	–	25.359
Probit-Linear	–	–	19.671	27.347
KHB Probit	–	–	19.697	27.467
Y-standardization A: Probit	–	–	24.490	30.897
Y-standardization B: Probit	–	29.556	19.697	33.760
IPTW	–	30.609	–	26.944

Values represent the percent mediated (the ratio of the indirect effect to the total effect), using various combinations of mediator and outcome variable types

See Table 2 for description of methods

logit), and it returns estimates of effects on the latent scale identified up to scale; that is, it returns probit estimates (on the scale defined by the probit model including all variables). Because these effects are not comparable with the effects on the probability margin reported thus far, we make use of the result that the KHB method also applies to average partial effects (as defined in Wooldridge 2002). The estimate of the direct average partial effect lies between the estimates of Baron and Kenny (1986) and Imai et al. (2010a, b, c) (−0.050), but is otherwise similar, and the estimate of the indirect average partial effect is identical to those previously reported (−0.012).

In the final two rows in Table 2 we report the results using Y-standardization A—applying the "product of standardized coefficients" method—and B—applying the "difference of standardized coefficients" method (MacKinnon and Dwyer 1993). Not only do these methods return effects on scales different from each other, their scales also differ from the remaining approaches. Y-standardization A reports how a standard deviation unit increase in the treatment variable changes the scale of latent Y measured in standard deviations, while Y-standardization B reports the treatment effect on the scale of latent Y measured in standard deviations. For example, the results using Y-standardization B suggest that the total treatment effect is −0.188 standard deviations on the latent scale of Y, while the indirect effect is −0.046. In the scenario with a binary treatment variable, Y standardization B thus appears to be more meaningful than Y-standardization A.

4.1.2 Mediation percentages

So far we have given interpretation to estimates of total, direct, and indirect effects. In this section we present results in terms of the mediation percentage, i.e., the ratio of the indirect effect to the total effect. In Table 3 we compare these mediation percentages estimates across methods and across the four situations previously defined. In the first column, we consider the case in which both outcome and mediator are continuous. According to Hicks and Tingley (2011), the results should be similar between the linear SEM approach and the approach by Imai et al. (2010a, b, c), but the percentages differ by ∼5 percentage points (27 vs 22 %). Since the point estimates reported in Table 1 are virtually identical, this

difference is likely the result of the simulation-based approach of Imai et al. (2010a, b, c) (see footnote 10). In this approach, the mediation percentage is calculated for each repetition in the simulation study, yielding a distribution of mediation percentages. The reported mediation percentage is, in this setup, the median of this percentage distribution.

In the second column—continuous outcome, binary mediator—we find overall agreement of roughly 30 % across methods except for the linear SEM approach, which, similar to the first situation, returns a considerably higher mediation percentage (39 %). However, the two models using the approach by Imai et al. (2010a, b, c) differ by roughly 5 % points, indicating that choice of link functions for the outcome and mediator models is not arbitrary.

In the third column in Table 3 we investigate the situation with a binary outcome and a continuous mediator. Similar to the previous scenario, we find overall agreement across methods, except for the Y-standardization A approach. We also find that results based on the linear SEM, using a linear probability model for the outcome, and the equivalent method by Imai et al. (2010a, b, c) return near-identical results. The KHB method, which uses the "product of coefficient" method in probit model for the outcome of a linear model for the mediator, and the equivalent method by Imai et al. (2010a, b, c) also return near-identical results.

In the final column of Table 3 we examine the situation where both outcome and mediator are binary. We once again find overall agreement between methods, although both Y-standardization approaches return estimates of the mediation percentages above the other methods. The pattern of results is also similar to the previous situation in which similar link functions return similar results.

In summary, despite the differing motivations and formulations behind the mediation analysis methods applied to the JOBS II dataset, the methods appear to return very similar results across the four scenarios. The exception appears to be the method of Y-standardization, which returns higher estimates of the mediation percentages than the other approaches in the binary cases, a result similar to the one found in the Monte Carlo simulations reported in Karlson et al. (2012).

4.2 A Monte Carlo study

In our second comparison of mediation approaches we conducted an extensive Monte Carlo simulation study to examine a situation which is often encountered in health services research—when the outcome, treatment, and mediator are all binary. As in the second JOBS II example, we focus on the extent to which the mediator mediates the treatment effect on the outcome using mediation percentages, i.e., the ratio of the indirect effect to the total effect. Our study was based on the following model:

$$\begin{aligned} T &= I(r < T^* = C + v) \\ M &= I(0 < M^* = \theta T + C + u) \\ Y &= I(q < Y^* = \beta T + \gamma M + C + e) \end{aligned}$$

where $I(.)$ is an indicator function, taking the value 1 when condition met, 0 when not met. C is a continuous confounder, T is a binary variable with threshold r chosen to yield distributions 30/70, 50/50, or 70/30 in three different simulations, respectively, M is a binary variable, v and u are drawn from standard normal distributions, e is drawn from a standard logistic distribution, and the threshold, q, is chosen such that Y takes on the following distributions: 50/50, 75/25, 95/5. Furthermore, we vary the magnitude of θ

Table 4 Monte Carlo simulation study of mediation approaches when treatment, mediator, and outcome are binary

		$N = 200$[a]		$N = 750$		$N = 2{,}500$	
		Mean absolute bias	Median absolute bias	Mean absolute bias	Median absolute bias	Mean absolute bias	Median absolute bias
A	Linear SEM (linear probability model)	14.417	7.097	50.240	11.742	57.364	9.757
B	Imai et al. (2010a, b, c): Logit-Logit	9.908	5.300	9.146	5.654	7.923	5.085
C	Imai et al. (2010a, b, c): Logit-Linear	9.462	5.208	8.337	6.714	6.316	4.050
D	KHB Logit	10.477	5.030	8.961	6.483	6.502	4.060
E	Y-standardization A	11.415	5.732	14.055	10.900	11.337	9.914
F	Y-standardization B	11.830	6.688	9.385	6.697	6.912	5.142
G	IPTW Logit	14.008	8.252	20.834	8.194	13.543	6.934
H	IPTW Stabilized Logit	11.768	9.364	16.050	8.748	15.777	7.503

Mean absolute bias refers to absolute deviation from the true percent mediated measured in percentage points averaged over 72 scenarios (with the true percent mediated ranging from 0 to 50 %). Median absolute bias reports the median of the 72 absolute deviations, measured in percentage points. True outcome and mediator models are logistic. 250 replications

The true percent mediated is defined as the ratio of the indirect effect to the total effect obtained from a Monte Carlo study using 100 replications and 1,000,000 observations per draw. Simulation setup available upon request. See description of methods in notes to Table 2

[a] As a result of two few observations in scenarios involving a 95/5-distribution of the binary outcome, we report means and medians of 48 scenarios for N = 200

across four values to obtain different correlations between x and z. Finally, we use two combinations of β and γ: $\beta = 1$ and $\gamma = 0.5$, and $\beta = 0.5$ and $\gamma = 1$. The setup yields 72 different scenarios (with the true percent mediated ranging from 0 to 50 %).[15] Our study is based on 250 replicates using 200, 750, and 2,500 observations per draw, respectively. Table 4 summarizes the results (simulation output is available upon request). It reports for each method the mean and median of the absolute bias—defined as the absolute deviation from the true percent mediated—over the 72 scenarios.[16]

Row A in Table 4 shows that the linear SEM approach using a linear probability model has a large bias across all sample sizes; a result which is consistent with that reported in Karlson et al. (2012). Further inspection of the simulations suggests that this bias arises when the binary outcome variable has a 95/5-distribution, suggesting that the linear probability model fails when the outcome is highly skewed. Rows B, C, and D show the respective biases for the method by Imai et al. (Imai et al. 2010a, b, c), which uses a logit link for both the outcome and mediator, the method using a logit link for the outcome and a linear model for the mediator, and the method by Karlson et al. (2012). These methods

[15] Because the true mediation effect cannot be analytically derived in this setup, we obtain the true percent mediated using a Monte Carlo study with 100 replications and 1,000,000 observations per draw, which essentially provides us with a population estimate.

[16] We report both mean and median given the skewed distribution of the mediation percentages across the 72 scenarios. Although the level of bias differs between the two central tendency measures, the overall pattern of results is very similar whether one uses the mean or the median as the basis of evaluation. We nevertheless report both central tendencies in order for the reader to properly assess the results.

return the lowest biases among all methods across all sample sizes.[17] In rows E and F, we report the biases of the approaches using the method of Y-standardization. Y-standardization A—the product of standardized coefficients—returns the largest biases. Y-standardization B—the difference in standardized coefficients—performs much better, returning the fourth lowest bias of all methods. Nevertheless, in their Monte Carlo study, Karlson et al. (2012) found that Y-standardization B failed in recovering the true mediation percentage in situations where the distribution of the mediator is very different from the error in the latent linear model underlying the logit model; a scenario not explored in the simulations we carried out here. In the final two rows, G and H, we report the results for the inverse probability of treatment weighting approach. For both methods, we find quite substantial biases across all sample sizes.

Our Monte Carlo study suggests that some methods perform better than others in recovering the true percent mediated when treatment, mediator, and outcome are all binary. We find that the methods by Imai et al. (2010a, b, c) and the method by Karlson et al. (2012) return, on average, the lowest absolute bias among all methods. Interestingly, the two specifications of the method by Imai et al. (2010a, b, c) return quite similar results, suggesting that using the linear model for a binary mediator works as well as using a (nonlinear) logit link. Our study also shows that Y-standardization B performs well, but given the results reported in Karlson et al. (2012), we suggest that researchers take care in employing this method. The remaining methods perform less satisfactorily: Y-standardization A and the IPTW methods return large biases. Perhaps most strikingly, using the linear SEM (a linear probability model) for meditational analysis returns biased results when outcome and mediator are binary, and we consequently recommend that researchers do not use this method.

4.3 Summary of approach comparison

In analyzing the JOBS II data, we found that point estimates of total, direct, and indirect effects were quite similar across methods for both continuous and binary outcomes, while methods appeared to disagree more on the reported percent mediated. Because these results are difficult to evaluate with observational data, we used simulated data to compare approaches against a common baseline. Taken together, our comparison of approaches using both real and simulated data suggests, first, that some methods perform better than others and, second, that those methods that perform best are very similar in their performance. The approach by Imai et al. (2010a, b, c) is among the best performers overall, is directly formulated in the potential outcomes framework, and—given its versatility in terms of models for outcomes and mediators—can be applied to most scenarios often met in applied research. While these characteristics speak to the advantages of using the approach of Imai et al. (2010a, b, c), we find that non-simulation based approaches appear to work just as well in terms of recovering mediation. For continuous outcomes, the method by Imai et al. (2010a, b, c) appears to default to the standard linear SEM approach, and for binary outcomes, the approach by Imai et al. (2010a, b, c) yields results highly similar to those obtained by the method by Karlson et al. (2012). Our study also suggested that the linear SEM approach (the linear probability model) appears to yield biased estimates when the binary outcome is highly skewed (and the linear approximation no longer holds), thereby supporting the contention that researchers should not use this approach for non-linear modeling.

[17] Comparing the method by Imai et al. (2010a, b, c) using a logit link for the outcome and a linear model for the mediator and the method by Karlson et al. (2012) in a Monte Carlo study, Breen et al. (Forthcoming) found the methods to yield highly similar results; corroborating the results we report here.

5 Discussion

In this paper, we sought to achieve two aims: to make the case for broader use of mediation analysis to better understand casual pathways in DM interventions and to provide a detailed discussion of the range of available approaches to conduct mediation analysis under different scenarios (e.g., a continuous versus dichotomous outcome or mediator). Both of these are relevant to other evaluations of large scale healthcare interventions as well. Like DM, evaluations of healthcare interventions typically focus on whether treatment effects were achieved and rarely explore the theorized underlying causal mechanism. This situation is problematic when the intervention is found not to work because we have limited insight into why the intervention failed so that it can be redesigned accordingly. More broadly, mediation analysis helps us better understand the nature and extent of underlying casual mechanisms, which can inform the design of related interventions. While we feel that there is a broader role for mediation analysis, it does not replace the critical role of experts who, at the outset, use their knowledge of the phenomenon to design the intervention, identify potential mediators, and assist in the application of mediation analysis through the selection of appropriate models and validation of assumptions. Then, once sufficient data is collected on the relevant program elements (i.e., treatment condition, baseline covariates, mediator and outcome), mediation analysis can be used to test whether the hypothesized causal mechanisms operate as expected. Referring back to the example given in the Introduction, if patient self-efficacy is on the causal pathway between a DM intervention in which nurses engage with patients by phone to promote healthier behaviors and the target outcomes of DM programs—fewer hospitalizations and reduced health care costs, the treatment group should experience a larger increase in self-efficacy than the control group, and increased self-efficacy should also lead to decreased costs. If the evaluation confirms these relationships, purchasers can feel more confident that their investment in these services will be rewarded with lower health care spending, program administrators can feel more confident that their intervention is operating effectively along a specified causal pathway, and behavioral change experts gain further support for their theory.

Equally important, mediation analysis can be informative in the absence of a treatment effect in three ways: (1) if the intervention increases self-efficacy, but increased self-efficacy does not lead to reduced health care costs, then such a finding points to the need to refine the theory on self-efficacy and its association with cost outcomes. It also suggests that other potential mechanisms should be considered for inclusion in the intervention, such as patient "activation" (Hibbard et al. 2004), psychological "sense of control" (Mirowsky and Ross 1991), or "self-care agency" (Sousa et al. 2010); (2) if the intervention does not increase self-efficacy, but increased self-efficacy that occurs on its own is found to be associated with lower costs, then the intervention requires refinement. Nurses' competency in improving self-efficacy should be investigated, as well as other potential issues limiting the effective delivery of the intervention (Butterworth et al. 2007); (3), the intervention does not increase naturally-varying self-efficacy, and differences in self-efficacy are not associated with lower costs, then this result suggests that both the theory and the program design need to be revisited. As a result of mediation analysis, each of these scenarios offers distinct and helpful guidance on how to move towards an effective intervention—information that would not be produced by a typical evaluation that ends upon confirming the null hypothesis of no treatment effect.

Program evaluators are faced with fundamental issues when conducting mediation analysis, such as ensuring that all the important variables have been collected and correctly specified, choosing an analytic mediation framework appropriate for the given research

question, and interpreting the results in relation to the theoretical context. As the results of our analyses demonstrate, specifying the correct model relative to the variable type of the mediator and outcome is important and, as a consequence, researchers should be equipped to make an informed choice for the analysis at hand. We find that the framework by Imai et al. (2010a, b, c) provides a versatile approach to mediation analysis which has among the best performance in the analyses we conducted, but we also find that the non-simulation based approaches of the linear SEM for continuous outcomes and the method for binary outcomes by Karlson et al. (2012) has very similar performance. Nevertheless, given its generality, the method by Imai and colleagues extends to non-parametric models which might prove useful in many areas of research.

As discussed throughout this paper, there are several limitations to mediation analysis—mostly resulting from the strong untestable assumptions necessary to draw valid inferences about indirect effects. These issues are central in non-experimental studies, but they persist even in RCTs where randomization occurs only on initial treatment assignment, and not later at the level of the mediator. Therefore, mediation analyses should be considered exploratory until more scientifically rigorous studies can be conducted (Jo and Stuart 2012), such as those described in Imai et al. (2013). At the very least, sensitivity analyses should be conducted after non-experimental mediation analyses, in order to gauge the extent to which unobserved variables must confound the mediator-outcome relationship to change the interpretation of a mediation effect (Hafeman 2011; Imai et al. 2010a, b, c; Jo and Vinokur 2011; VanderWeele 2010). Mediation effects that appear insensitive to unobserved variables increase our confidence in validity of the results, although sensitivity analysis in and of itself relies on untestable assumptions. Perhaps simply replicating the results using different methods might be a good strategy. In fact, it is probably the easiest to do and understand, and could be the most compelling confirmation of our temporary conclusion about mediation, which is based on strong untestable assumptions. Most importantly, these issues suggest that policy decisions should not be based on results from causal mediation analyses alone, but should rather be informed, first and foremost, by robust analyses using experimental designs.

While we have described many of the salient issues in mediation analysis, there are many features that remain beyond the scope of the current article. For example, Krull and MacKinnon (2001), Mathieu and Taylor (2007), and Zhang et al. (2009) describe the use and limitations of hierarchical (or mixed-model) approaches to test for multi-level mediation. Additionally, several papers have described the use and limitations of various longitudinal modeling strategies to test mediation effects (Bauer et al. 2006; Cheong et al. 2003; Cole and Maxwell 2003; Maxwell and Cole 2007; Maxwell et al. 2011; Selig and Preacher 2009; MacKinnon 2008, Chap. 8). Moreover, the application of instrumental variable techniques to mediation analysis is gaining in popularity amongst researchers outside of the field of economics (Antonakis et al. 2010; Gennetian et al. 2008; Sobel 2008). Finally, as briefly mentioned before, several approaches to sensitivity analysis have been developed for mediation studies, and should be considered as an integral post-estimation component of any mediation analysis (Hafeman 2011; Imai et al. 2010a, b, c; Jo et al. 2011; VanderWeele 2010). We encourage readers seeking to further broaden their understanding of mediation analysis to refer to these as well as other areas.

6 Conclusion

Over the past two decades, large-scale DM programs have repeatedly failed to deliver anticipated cost savings. In order for DM to be a viable strategy for reducing health care

costs in the future, the basic components of their intervention need to be examined in a way that can identify the cause of the failure. In this paper, we have described in detail the various methods available to systematically test hypothesized causal pathways so that effective interventions can be developed. After testing several competing models using real and simulated data, we find that some, but not all, models produce comparable results, once the mediator and outcome variable types are matched with the appropriate modeling strategy. We recommend that existing and future DM interventions be designed in a manner that allows for the regular testing of causal mechanisms in an effort to improve the likelihood of achieving desired outcomes.

Acknowledgments We thank Dustin Tingley, Raymond Hicks, Danella Hafeman, and Adam Glynn for clarifications of the modeling approaches used in their respective papers, to John Antonakis for evocative discussions pertaining to concerns of endogeneity in mediation analysis, and to Julia Adler-Milstein for her invaluable review and edits. We are indebted to the editor and two anonymous reviewers for providing excellent comments which substantially improved the manuscript.

References

Alwin, D.F., Hauser, R.M.: The decomposition of effects in path analysis. Am. Sociol. Rev. **40**, 37–47 (1975)

Antonakis, J., Bendahan, S., Jacquart, P., Lalive, R.: On making causal claims: a review and recommendations. Leadersh. Q. **21**, 1086–1120 (2010)

Baron, R.M., Kenny, D.A.: The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol. **51**, 118–1173 (1986)

Bauer, D.J., Preacher, K.J., Gil, K.M.: Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. Psychol. Methods **11**, 142–163 (2006)

Bodenheimer, T., Lorig, K., Holman, H., Grumbach, K.: Patient self-management of chronic disease in primary care. J. Am. Med. Assoc. **288**, 2469–2475 (2002)

Bollen, K.A.: Structural equations with latent variables. Wiley, New York (1989)

Breen, R.B., Karlson, K.B., Holm, A.: Total, direct, and indirect in logit and probit models. Sociol. Methods Res. (Forthcoming)

Butterworth, S.W., Andersen, B.T.: Health Coaching Performance Assessment™ (HCPA): a new tool for benchmarking and improving effectiveness. HealthSciences Institute. http://healthsciences.org/health-coaching-performance-assessment-hcpa-white-paper(2011). Accessed 13 Feb 2012

Butterworth, S., Linden, A., McClay, W.: Health coaching as an intervention in health management programs. Dis. Manag. Health Outcomes **15**, 299–307 (2007)

Cheong, J., MacKinnon, D.P., Khoo, S.T.: Investigation of meditational process using parallel process latent growth curve modeling. Struct. Equ. Model. **10**, 238–262 (2003)

Cole, D.A., Maxwell, S.E.: Testing meditational models with longitudinal data: questions and tips in the use of structural equation modeling. J. Abnorm. Psychol. **112**, 558–577 (2003)

Congressional Budget Office: an analysis of the literature on disease management programs. Washington DC: Congressional Budget Office. http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/59xx/doc5909/10-13-diseasemngmnt.pdf(2004). Accessed 19 Oct 2012

Cramer, J.S.: Logit models. From economics and other fields. Cambridge University Press, Cambridge (2003)

Duncan, O.D.: Path analysis: sociological examples. Am. J. Sociol. **72**, 1–16 (1966)

Efron, B., Tibshirani, R.: An introduction to the bootstrap. Chapman and Hall, New York (1993)

Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. Biometrics **58**, 21–29 (2002)

Freedman, L.S., Schatzkin, A.: Sample size for studying intermediate endpoints within intervention trials of observational studies. Am. J. Epidemiol. **136**, 1148–1159 (1992)

Gennetian, L.A., Magnuson, K., Morris, P.A.: From statistical associations to causation: what developmentalists can learn from instrumental variables techniques coupled with experimental data. Dev. Psychol. **44**, 381–394 (2008)

Glynn, A.N.: The product and difference fallacies for indirect effects. Am. J. Political Sci. **56**, 257–269 (2012)

Goetzel, R.Z., Ozminkowski, R.J., Villagra, V.G., Duffy, J.: Return on investment on disease management: a review. Health Care Financ. Rev. **26**, 1–19 (2005)
Hafeman, D.M.: Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. Am. J. Epidemiol. **174**, 710–717 (2011)
Hafeman, D.M., Schwartz, S.: Opening the black box: a motivation for the assessment of mediation. Int. J. Epidemiol. **38**, 838–845 (2009)
Hibbard, J.H., Stockard, J., Mahoney, E.R., Tusler, M.: Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv. Res. **39**, 1026–1105 (2004)
Hicks, R., Tingley, D.: Casual mediation analysis. Stata J. **11**, 605–619 (2011)
Hill, J., Waldfogel, J., Brooks-Gunn, J.: Sustained effects of high participation in an early intervention for low-birth-weight premature infants. Dev. Psychol. **39**, 730–744 (2003)
Hirano, K., Imbens, G.W.: The propensity score with continuous treatments. In: Gelman, A., Meng, X.-L. (eds.) Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, pp. 73–84. Wiley InterScience, West Sussex (2004)
Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**, 945–960 (1986)
Holland, P.W.: Causal inference, path analysis, and recursive structural equation models. In: Clogg, C.C. (ed.) Sociological Methodology, pp. 449–484. American Sociological Association, Washington, DC (1988)
Hong, G.: Ratio of mediator probability weighting for estimating natural direct and indirect effects. In: 2010 Proceedings of the American Statistical Association, Biometrics Section, pp. 2401–2415. American Statistical Association, Alexandria (2010)
Imai, K., Keele, L., Tingley, D.: A general approach to causal mediation analysis. Psychol. Methods **15**, 309–334 (2010a)
Imai, K., Keele, L., Yamamoto, T.: Identification, inference, and sensitivity analysis for causal mediation effects. Stat. Sci. **25**, 51–71 (2010b)
Imai, K., Keele, L., Tingley, D., Yamamoto, T.: Advances in social science research using R. In: Vinod, H.D. (ed.) Causal Mediation Analysis Using R, pp. 129–154. Springer, New York (2010c)
Imai, K., Tingley, D., Yamamoto, T.: Experimental designs for identifying causal mechanisms. J. R. Stat. Soc. A **176**(1), 5–51 (2013)
Imai, K., van Dyke, D.A.: Causal inference with general treatment regimes: generalizing the propensity score. J. Am. Stat. Assoc. **99**, 854–866 (2004)
Jo, B.: Causal inference in randomized experiments with mediational processes. Psychol. Methods **13**, 314–336 (2008)
Jo, B., Stuart, E.A.: Comments: causal interpretations of mediation effects. J. Res. Educ. Eff. **5**, 250–253 (2012)
Jo, B., Stuart, E.A., MacKinnon, D.P., Vinokur, A.D.: The use of propensity scores in mediation analysis. Multivar. Behav. Res. **46**, 425–452 (2011)
Jo, B., Vinokur, A.D.: Sensitivity analysis and bounding of causal effects with alternative identifying assumptions. J. Educ. Behav. Stat. **36**, 415–440 (2011)
Joffe, M.M., Rosenbaum, P.R.: Invited commentary: propensity scores. Am. J. Epidemiol. **150**, 327–333 (1999)
Judd, C.M., Kenny, D.A.: Process analysis: estimating mediation in treatment evaluations. Eval. Rev. **5**, 602–619 (1981)
Karlson, K.B., Holm, A.: Decomposing primary and secondary effects: a new decomposition method. Res. Stratif. Soc. Mobil. **29**, 221–237 (2011)
Karlson, K.B., Holm, A., Breen, R.: Comparing regression coefficients between models using logit and probit: a new method. Sociol. Methodol. **42**, 274–301 (2012)
Kohler, U., Karlson, K.B., Holm, A.: Comparing coefficients of nested nonlinear probability models. Stata J. **11**, 420–438 (2011)
Kraemer, H.C., Kiernan, M., Essex, M.J., Kupfer, D.J.: How and why criteria defining moderators and mediators differ between the Baron and Kenny and MacArthur approaches. Health Psychol. **27**, 101–108 (2008)
Krull, J.L., MacKinnon, D.P.: Multilevel modeling of individual and group level mediated effects. Multivar. Behav. Res. **36**, 249–277 (2001)
Linden, A., Adler-Milstein, J.: Medicare disease management in a policy context. Health Care Financ. Rev. **29**, 1–11 (2008)
Linden, A., Adams, J.L.: Using propensity score-based weighting in the evaluation of health management programme effectiveness. J. Eval. Clin. Pract. **16**, 175–179 (2010a)

Linden, A., Adams, J.L.: Evaluating health management programmes over time: application of propensity score-based weighting to longitudinal data. J. Eval. Clin. Pract. **16**, 180–185 (2010b)
Linden, A., Roberts, N.: Disease management interventions: what's in the black box? Dis. Manag. **7**, 275–291 (2004)
Linden, A., Butterworth, S., Roberts, N.: Disease management interventions II: what else is in the black box? Dis. Manag. **9**, 73–85 (2006)
Long, J.S.: Regression models for categorical and limited dependent variables. Sage, Thousand Oaks (1997)
Lorig, K.R., Holman, H.: Self-management education: history, definition, outcomes, and mechanisms. Ann. Behav. Med. **26**, 1–7 (2003)
Lu, B., Zanutto, E., Hornik, R., Rosenbaum, P.R.: Matching with doses in an observational study of a media campaign against drug abuse. J. Am. Stat. Assoc. **96**, 1245–1253 (2001)
MacKinnon, D.P.: Introduction to Statistical Mediation Analysis. Erlbaum, Mahwah, NJ (2008)
MacKinnon, D.P., Dwyer, J.H.: Estimation of mediated effects in prevention studies. Eval. Rev. **17**, 144–158 (1993)
MacKinnon, D.P., Warsi, G., Dwyer, J.H.: A simulation study of mediated effect measures. Multivar. Behav. Res. **30**, 41–62 (1995)
MacKinnon, D.P., Lockwood, C.M., Brown, C.H., Wang, W., Hoffman, J.M.: The intermediate endpoint effect in logistic and probit regression. Clin. Trials **4**, 499–513 (2007)
MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., Sheets, V.: A comparison of methods to test mediation and other intervening variable effects. Psychol. Methods **7**, 83–104 (2002)
Manski, C.F.: Identification of treatment response with social interactions. Econ. J. (Forthcoming)
Marks, R., Allegrante, J.P., Lorig, K.L.: A review and synthesis of research evidence for self-efficacy-enhancing interventions for reducing chronic disability: implications for health education practice (Part I). Health Promot. Pract. **6**, 37–43 (2005)
Matheson, D., Wilkins, A., Psacharopoulos, D.: Realizing the promise of disease management: payer trends and opportunities in the United States. Boston Consulting Group, Boston (2006)
Mathieu, J.E., Taylor, S.R.: A framework for testing meso-mediational relationships in organizational behavior. J. Organ. Behav. **28**, 141–172 (2007)
Mattke, S., Seid, M., Ma, S.: Evidence for the effect of disease management: is $1 billion a year a good investment? Am. J. Manag. Care **13**, 670–676 (2007)
Maxwell, S.E., Cole, D.A.: Bias in cross-sectional analyses of longitudinal mediation. Psychol. Methods **12**, 23–44 (2007)
Maxwell, S.E., Cole, D.A., Mitchell, M.A.: Bias in cross-sectional analyses of longitudinal mediation: partial and complete mediation under an autoregressive model. Multivar. Behav. Res. **46**, 816–841 (2011)
Mays, G.P., Au, M., Claxton, G.: Convergence and dissonance: evolution in private-sector approaches to disease management and care coordination. Health Aff. **26**, 1683–1691 (2007)
McKelvey, R.D., Zavoina, W.: A statistical model for the analysis of ordinal level dependent variables. J. Math. Sociol. **4**, 103–120 (1975)
Miller, W.R., Rose, G.S.: Toward a theory of motivational interviewing. Am. Psychol. **64**, 527–537 (2009)
Mirowsky, J., Ross, C.E.: Eliminating defense and agreement bias from measures of the sense of control: a 2×2 index. Soc. Psychol. Q. **54**, 127–145 (1991)
Morgan, S.L., Todd, J.J.: A diagnostic routine for the detection of consequential heterogeneity of causal effects. Sociol. Methodol. **38**, 231–281 (2008)
Nelson, L.: Lessons from medicare's demonstration projects on disease management and care coordination. Congressional Budget Office Working Paper 2012-01. http://www.cbo.gov/ftpdocs/126xx/doc12664/WP2012-01_Nelson_Medicare_DMCC_Demonstrations.pdf.(2012). Accessed 11 Feb 2012
Ofman, J.J., Badamgarav, E., Henning, J.M., Knight, K., Gano Jr, A.D., Levan, R.K., Gur-Arie, S., Richards, M.S., Hasselblad, V., Weingarten, S.R.: Does disease management improve clinical and economic outcomes in patients with chronic diseases? A systematic review. Am. J. Med. **117**, 182–192 (2004)
Pearl, J.: Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. pp. 411–420. Morgan Kaufmann, San Francisco (2001)
Pearl, J.: The mediation formula: a guide to the assessment of causal pathways in non-linear models. Technical report R-363, University of California, Los Angeles (2011)
Pearl, J.: The causal foundations of structural equation modeling. In: Hoyle, R.H. (ed.) Handbook of Structural Equation Modeling, pp. 68–91. Guilford Press, New York (2012)
Peterson, M.L., Sinisi, S.E., van der Laan, M.J.: Estimation of direct causal effects. Epidemiology **17**, 276–284 (2006)
Robins, J.M.: Marginal structural models. In: 1997 Proceedings of the Section on Bayesian Statistical Science, pp. 1–10. American Statistical Association, Alexandria (1998)

Robins, J.M., Greenland, S.: Identifiability and exchangeability for direct and indirect effects. Epidemiology **3**, 143–155 (1992)
Robins, J.M., Hernan, M.A., Brumback, B.: Marginal structural models and causal inference in epidemiology. Epidemiology **11**, 550–560 (2000)
Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 41–55 (1983)
Royston, P., Altman, D.G., Sauerbrei, W.: Dichotomizing continuous predictors in multiple regression: a bad idea. Stat. Med. **25**, 127–141 (2006)
Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
Rubin, D.B.: Bayesian inference for causal effects: the role of randomization. Ann Stat **6**, 34–58 (1978)
Shrout, P., Bolger, N.: Mediation in experimental and nonexperimental studies: new procedures and recommendations. Psychol. Methods **7**, 422–445 (2002)
Selig, J.P., Preacher, K.J.: Mediation models for longitudinal data in developmental research. Res. Hum. Dev. **6**, 144–164 (2009)
Sobel, M.E.: Asymptotic confidence intervals for indirect effects in structural equation models. In: Leinhardt, S. (ed.) Sociological Methodology, pp. 290–312. American Sociological Association, Washington, DC (1982)
Sobel, M.E.: Identification of causal parameters in randomized studies with mediating variables. J. Educ. Behav. Stat. **33**, 230–251 (2008)
Sousa, V.D., Zauszniewski, J.A., Bergquist-Beringer, S., Musil, C.M., Neese, J.B., Jaber, A.F.: Reliability, validity and factor structure of the Appraisal of Self-Care Agency Scale—Revised (ASAS-R). J. Eval. Clin. Pract. **16**, 1031–1040 (2010)
Stolzenberg, R.M.: The measurement and decomposition of causal effects in nonlinear and nonadditive models. Sociol. Methodol. **11**, 459–488 (1980)
VanderWeele, T.J.: Marginal structural models for the estimation of direct and indirect effects. Epidemiology **20**, 18–26 (2009)
VanderWeele, T.J.: Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology **21**, 540–551 (2010)
Winship, C., Mare, R.D.: Structural equations and path analysis for discrete data. Am. J. Sociol. **89**, 54–110 (1983)
Winship, C., Mare, R.D.: Regression models with ordinal variables. Am. Sociol. Rev. **49**, 512–525 (1984)
Wooldridge, J.M.: Econometric analysis of cross section and panel data. MIT Press, Cambridge (2002)
Zanutto, E., Lu, B., Hornik, R.: Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. J. Educ. Behav. Stat. **30**, 59–73 (2005)
Zhang, Z., Zyphur, M.J., Preacher, K.J.: Testing multilevel mediation using hierarchical linear models: problems and solutions. Organ. Res. Methods **12**, 695–719 (2009)