Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
21(1) 77–107

DOI: 10.1177/0962280210391076
smm.sagepub.com
SAGE

A review of causal estimation of effects in mediation analyses

Thomas R Ten Have and Marshall M Joffe

Abstract

We describe causal mediation methods for analysing the mechanistic factors through which interventions act on outcomes. A number of different mediation approaches have been presented in the biomedical, social science and statistical literature with an emphasis on different aspects of mediation. We review the different sets of assumptions that allow identification and estimation of effects in the simple case of a single intervention, a temporally subsequent mediator and outcome. These assumptions include various no confounding assumptions including sequential ignorability assumptions and also interaction assumptions involving the treatment and mediator. The understanding of such assumptions is crucial since some can be assessed under certain conditions (e.g. treatment–mediator interactions), whereas others cannot (sequential ignorability). These issues become more complex with multiple mediators and longitudinal outcomes. In addressing these assumptions, we review several causal approaches to mediation analyses.

Keywords

baseline randomisation, direct effects, principal stratification, sequential ignorability, structural mean models, treatment–mediator interactions, unmeasured confounding

1 Introduction

With intervention studies or risk factor exposure investigations in biomedical research, researchers want to know if an intervention has a direct effect on outcome that involves certain unmeasured post-baseline factors not accounted for by measured post-baseline factors.[1,2] As in one of our examples, such measured post-baseline factors may represent undesirable effects on outcome (e.g. depression). There is interest in understanding if the intervention (e.g. specific therapy-based intervention) has a direct effect on the outcome involving unmeasured factors that are beneficial for the outcome, rather than measured factors that are not beneficial in certain instances (e.g. other therapies). Alternatively, interest may focus on direct effects of an intervention (e.g. BHS, behavioural health specialist) involving unmeasured post-randomisation factors beyond the well-established positive outcome benefits of measured post-baseline factors (e.g. medication). In these contexts, we review standard and causal mediation approaches for direct and, to a lesser extent,

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA.

Corresponding author:
Thomas R Ten Have, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA
Email: ttenhave@upenn.edu

indirect effects that require various combinations of no unmeasured confounding (ignorability) and intervention homogeneity (no intervention–mediator interactions).

The specific clinical context for this review is a clinical trial of a cognitive behavioural intervention (CBT, Cognitive behavioural theropy) *versus* usual care with respect to depression in patients who had recently attempted suicide. The cognitive behavioural intervention focused on short-term resolutions of dysfunctional psychological and cognitive function and behaviours through a goal-oriented, systematic procedure. One of the goals of this intervention was to reduce reliance on other therapies that often sought out by suicide attempters. Such therapies are typically of longer term than CBT and thus may not address the acute needs (e.g. reducing depression) of suicide attempters as quickly as CBT. Consequently, seeking of non-study therapies was measured as a possible mediator. For the general discussion, we refer to the intervention (e.g. CBT), which is randomised at baseline, as the randomised intervention. The group randomised to the baseline intervention is the randomised intervention group, and the group randomised to the non-intervention group (e.g. usual care) is the randomised comparison group. The mediator (e.g. external non-CBT) is the post-randomisation factor measured for all study participants and occurring after the randomised intervention is assigned but not necessarily finished. The outcome (e.g. depression) is the measured response variable occurring after the randomised intervention and mediator. The pathway between the randomised baseline intervention and the outcome through this mediation variable is the indirect effect, and the pathway between the randomised baseline intervention and outcome bypassing this mediator is the direct effect. We focus on the direct effect for the reasons described above.

While this article focuses on the above CBT study for illustrating the reviewed methods, the data analysis section includes a suicide prevention study. It compared collaborative care management for treating depression (and thus reducing the risk of suicide) with usual care in 293 elderly depressed primary care patients.[3] The collaborative care management programme in the intervention group was based on patient, primary care, staff and physician interactions with a nurse-level BHS. We evaluate if the significant intent-to-treat (ITT) effect of the intervention on the 4-month Hamilton depression outcome was due to a direct effect apart from the mediator, as described above, use of prescriptive anti-depressant medication between baseline and 4 months.

Pearl[4] and Robins[5] defined the 'controlled' or 'prescriptive' direct effect in that it requires that the intervention assignment as well as the potentially uncontrollable mediator factor (e.g. external non-CBT) be manipulated by a clinical investigator to a pre-specified value. Without stating such, standard mediation methods going back to Baron and Kenny[6] have focused on controlled direct effects.[7] In addition, these methods have addressed indirect effects, but such effects are not interpretable except under the assumption of no interaction.

The interpretation of the controlled direct effect as being manipulable by a clinician to a fixed level may be implausible for hard-to-control behavioural or clinical mediation factors such as pain interference and intermediate risk factors for depression. In contrast, if the mediator represents another intervention impacted by the randomised intervention then setting this second intervention to a particular level has clinical significance and it is plausible that one could have a strategy for manipulating this second intervention regardless of the status of the first intervention assignment.

Robins,[5] Pearl[4] and others have proposed natural direct effects that would be 'realised if the 'randomised treatment' were administered, but its effect on the mediator were somehow blocked, or equivalently, if the mediator were kept at the level it would have taken in the absence of the 'randomised treatment'.[7] A similar interpretation can be made for the natural indirect effect but

blocking the randomised intervention rather than the mediator. Under linear models without individual-level interactions, the natural direct effect is equal to the controlled direct effect.[8] This is not true for linear models with interactions between the intervention and mediator as well as non-linear models such as the logistic model.[7]

The above controlled direct and indirect causal effects can be specified as contrasts between potential or counterfactual outcomes under the general class of causal models known as the 'Rubin Causal Model' (RCM).[9,10] Under the RCM, causal inference is defined in terms of contrasts among multiple prospective or potential outcomes defined under different conditions for the same individual, holding all other factors, observed and unobserved, constant.[11] In general, because each potential outcome represents a unique pair of intervention and mediator level combinations, it cannot be observed simultaneously with another potential outcome.[5] For example, when the intervention is CBT and the mediator is seeking external non-study therapy, four pairs of CBT/non-study therapy levels would be yes/yes, yes/no, no/yes and no/no. Each participant will only experience one of these four intervention/mediator pairs of levels.

Given this inability to simultaneously observe all potential outcomes and even to observe certain potential outcomes, standard mediation approaches (e.g. regression, path and structural equation model (SEM)) as well as recently proposed causal methods make different combinations of ignorability assumptions for the randomised intervention and the mediator (i.e. sequential ignorability; no unmeasured confounding) as well as no-interaction assumptions to identify controlled and natural direct and indirect effects. Ignorability assumptions are more likely to be satisfied with randomisation of both the intervention and mediator, but such randomisation does not occur with observational studies, and in most randomised studies only the intervention is randomised. In the context of CBT study, Ten Have et al.[12] presented contradicting evidence for the mediation of CBT on depression by non-study therapy depending on whether the sequential ignorability assumption was made. This example illustrates the practical vulnerabilities of methods and studies that assume sequential ignorability and thus the need for sensitivity analyses including methods that do not make this assumption.

In spite of the fact that ignorability of the mediator under sequential ignorability is unlikely in most randomised and non-randomised contexts, there is much literature on mediation approaches assuming sequential ignorability in a variety of contexts.[6,13–23] With a standard regression approach, Baron and Kenny[6] brought the concept of mediation to the forefront in psychological research. Their article suggested a sequence of steps for testing the mediation model. Recent study has extended this approach to more complex designs[19,20,23] and to settings involving multiple measurements of intervention level, mediators and outcomes over time.[22]

In response to the untenability of the sequential ignorability assumption even in a randomised trial context, some causal approaches[12,24] require that only the intervention be randomised but under alternative model assumptions. Such modelling assumptions relate to the absence of intervention–mediator interactions at the individual level,[5] which are required not only just for these causal approaches but also for standard mediation approaches under certain types of sequential ignorability (e.g. if post-baseline confounders are present). Two types of such interactions are addressed in the literature. First, there are causal interactions involving contrasts between potential outcomes.[5,25] Examples would be whether the controlled causal direct effect of the intervention differs across the specified levels of the mediator or whether the natural causal direct effect differs across the levels of the intervention. Second, there are interactions where the causal contrast for the mediator is constant across different groups of subjects with different intervention assignments (e.g. no-current treatment interaction.[26,27] That is, contrasts for pairs of potential outcomes are assumed to be constant for different observed groups of participants defined by the

observed randomised intervention and mediator. Subsequently in this article, we review these no-interaction and sequential ignorability assumptions in more detail.

The two causal modelling approaches reviewed here under the RCM framework in the randomised trial–mediation contexts are: (1) semi-parametric methods related to structural nested mean model (SNMM)[5,12,28] and (2) principal stratification (PS).[29–32] These two causal approaches represent very different mediation strategies, although they both make no assumptions about the mediator in terms of its relationships with measured and unmeasured confounders. The SNMM-based approach follows more closely the traditional regression method of Baron and Kenny[6] in terms of parameter interpretation. In contrast, the PS method stratifies the population into partially latent classes (principal strata) based on potential observations for the mediator variable under each of the levels of the randomised intervention. Mediation analyses are then based on ITT effects of the randomised intervention on outcome within selected principal strata, which VanderWeele[31] refers to as principal strata direct effects. Identification of these stratified ITT effects relies on relationships between baseline covariates and the probabilities of membership in these principal strata in addition to model assumptions for the outcome. Heterogeneity of ITT effects on outcome across these select principal strata provides one way of assessing the interactions involving the outcome as the dependent variable, with the caveat that comparing principal strata effects across principal strata is vulnerable to confounding. Additionally, interpretation of the PS effects does not require that an intervention exists or that it be conceivable that one could exist to control the mediator such that the controlled or natural effects are defined. Finally, the logistic model with binary outcomes under the PS approach is not vulnerable to problems with other causal mediation approaches resulting from the non-linearity of the logistic model.

We now address the specifies of the two behavioural intervention studies introduced above. They offer divergent conditions for illustrating the differences and similarities between the traditional and causal approaches. The first study is a suicide therapy study, which evaluated the effect of CBT versus usual care in treatment of suicide attempts, suicide ideation, hopelessness and depression in 120 patients who had recently attempted suicide.[33] The sample size for this investigation at 6 months is 101 due to drop-out, which appears to be weakly associated with the factors used in this analysis as well as others ($p > 0.35$).[33] We assess if the significant ITT effect of CBT on 6–month depression outcome as measured by the Beck Depression Inventory-II (BDI) was due to a direct effect apart from use of non-study therapy (mediator) between 4 and 6 months. Potential confounders of the mediator–outcome relationship include economic and personal stress reducing the motivation for non-study therapy and increasing the likelihood of depression in suicide attempters.

The second study, a suicide prevention study, compared collaborative care management for treating depression (and thus reducing the risk of suicide) with usual care in 293 elderly depressed primary care patients.[3] The collaborative care management programme in the intervention group was based on patient, primary care, staff and physician interactions with a nurse-level care manager. We evaluate if the significant ITT effect of the intervention on the 4-month Hamilton depression outcome was due to a direct effect apart from use of prescriptive anti-depressant medication (mediator) between baseline and 4 months. Potential unmeasured confounders of the medication–depression relationship include medical comorbidities at follow-up, which deter elderly depressed patients from taking anti-depressant medications because of so many other medications necessitated by their medical comorbidities, which also predispose patients to depression. As with the first study, potential baseline factors such as baseline depression and suicide ideation might have modified the significant effect of the care manager intervention and also the mediator, anti-depressant medication, on the follow-up depression outcome.

In the remainder of this article, we review different approaches to the identifiability and estimation of controlled and natural direct and indirect effects. Specifically, in Section 2, we start with the notation of potential outcomes that offer a strategy for presenting the assumptions for identifiability. We then present definitions of these causal effects in Section 3. Reviews of the assumptions and procedures that are used to identify and estimate these causal effects are subsequently presented for non-parametric estimation (Section 4), linear structural equation models (LSEM, Section 5), post-baseline confounders (Section 6), semi-parametric and PS cases (Section 7) and a variety of non-contexts representing departures from linear models (Section 8). Illustrations of the approaches with two applications are presented in Section 9. Finally, we summarise the presentation in Section 10.

2 Notation

In this section, we define notation for observed and potential outcome and mediator variables. Such notation is crucial in distinguishing between controlled and natural effects as well as the variety of assumptions and approaches that have been used to identify and estimate these effects.

2.1 Notation: observed random variables

First, we define the observed random variables, distinguishing them from the corresponding potential outcomes that would be observed under certain intervention and mediating factor conditions. Let Y denote the observed random variable for the outcome, which for this article is assumed to be continuous. We will briefly consider the case in Section 9 where the outcome, Y, is binary. Let R denote the observed binary random variable for the randomised intervention assignment such that $R=1$ if randomised to the randomised intervention; and $R=0$ if randomised to the comparison group. For most of the discussion, we assume the mediator, M, can be binary or continuous, and denote when one or other is needed for a specific method. With binary M, $M=1$ if the participant exhibits a positive level for the mediator (e.g. non-study therapy is not used); and $M=0$ if the participant does not exhibit a positive level for the mediator (e.g. non-study therapy is used).

To adjust for observed confounding, we also define notation for observed baseline and post-baseline confounders. Let $\mathbf{X}$ be the vector of observed baseline covariates, and $\mathbf{Z}$ be the vector of observed post-baseline confounders. The post-baseline confounders are measured before the mediator, M, is measured. The identifiability assumptions described below are distinguished in one way in terms of the presence or absence of the post-baseline confounders. Finally, we suppress the index i to simplify notation, but note here that all subsequent notation applies to the i-th of n participants.

2.2 Notation: potential random variables and counterfactuals

In defining the potential variable notation, we index the potential variables with a randomised intervention level, r, and the mediation level, m. The indices, r and m, are not necessarily the observed levels of the randomised intervention and mediation factors, but instead are specified or 'set' to define contrasts of the potential outcome variables for an individual participant.

Before proceeding to the potential outcome notation for the causal mediation models, we consider as an introduction the potential outcome notation for the simple ITT effect in a randomised trial. In this context, the RCM distinguishes between the observed outcome, Y and

the two potential outcomes denoted by Y_r ($r \in \{1, 0\}$), each of which would have been observed for that subject, had they been randomised to the comparison group or the intervention, respectively. One of these potential outcomes will be observed, while the other will be an unobserved, or counterfactual, outcome. The corresponding causal effect in this simple case is the ITT contrast between these two potential outcomes: i.e. $E[Y_1 - Y_0]$, which can be estimated in an unbiased way with the observed ITT difference between randomised intervention sample means. The PS approach specifies ITT contrasts within each principal stratum defined by potential mediation behaviour under each randomised intervention level.

We extend the potential outcomes framework to accommodate the mediation variable by using doubly indexed potential outcomes. Specifically, we let Y_{rm} denote the potential outcome for participant i that would occur if the randomised intervention, R, were set to level r, and if the mediator, M, were manipulated to level m. For the natural causal effects and the PS approach and when the mediator is thought not be manipulable, mediation inference is based on the potential mediator level under randomised intervention assignment r, as denoted by M_r. Furthermore, we define an additional potential outcome where we do not specify the actual level of the mediator, but allow it to be determined by the level of the randomised intervention assignment: $Y_{rM_{r'}}$, where r not necessarily is the same as r'.

Finally, in presenting the PS approach, we define the four principal strata for a binary mediator. Specifically, we define C as the PS indicator variable for the four principal strata: $C=1$, when $M_r=r$; $C=2$, when $M_1=M_0=1$; $C=3$, when $M_1=M_0=0$; and $C=4$ when $M_r=1-r$. The motivation for these principal strata definitions is provided in Section 3 in defining direct and indirect effects under the PS approach.

3 Causal effects

As mentioned in Section 1, the literature[1,4,5,8] has defined the following natural and controlled direct and indirect effects. The controlled or prescriptive direct effect for a fixed level of the mediator at m is:

$$\theta_{CDm} = Y_{1m} - Y_{0m} \tag{1}$$

for all m. Such an effect assumes that we can physically set the level of the mediator to a specific value. The analogous controlled effect of the mediator on the outcome is defined as:

$$\theta_{CMr} = Y_{rm} - Y_{rm'} \tag{2}$$

for $r=0$, 1 and all $m \neq m'$.

An alternative effect is the natural direct effect of the baseline intervention. The choice of a controlled or natural effect depends on the question. Under the CBT example, a controlled effect addresses the CBT effect on depression if a participant does not seek non-study therapy. The natural effect addresses the effect of CBT on depression if the intervention did not influence the decision to seek non-study therapy. The natural direct effect at a reference level set equal to r is:

$$\theta_{NDr} = Y_{1M_r} - Y_{0M_r} \tag{3}$$

for $r=0$, 1. Instead of setting the mediator variable, M, at a specific level m, we set M at the level that would be potentially achieved under assignment to level of r of the randomised intervention,

regardless of whether the participant was actually assigned to the randomised intervention or control group. When $r=0$ (the mediator potential variable is defined under the control assignment), the resulting effect can be interpreted as the effect of switching treatment level assignment for an individual on the outcome were the mediator to be held constant at its potential value under the control assignment, thus eliminating all indirect effects. Pearl[4] called θ_{ND0} a 'natural' direct effect, whereas Robins and Greenland[34] and Robins[5] named it a 'pure' direct effect. This natural or pure direct effect is the effect of the treatment that would be observed if a participant were assigned the mediator level that they would have had in the absence of treatment.[7] When $r=1$, Robins[5] called θ_{ND1} the 'total' direct effect. Regardless of the level to which r is set, we follow Imai et al.[8] by referring to θ_{NDr} as the natural direct effect. We similarly define the natural indirect or mediation effect at the treatment assignment level r:

$$\theta_{NIDr} = Y_{rM_1} - Y_{rM_0} \tag{4}$$

for $r=0, 1$. Here, the effect of the intervention on outcome is through it's impact on potential M, as quantified by changing the level of M from M_1 to M_0. Pearl[4] and Robins[5] called θ_{NID0} at $r=0$ the natural or pure indirect effect, respectively. Imai et al.[8] called θ_{NIDr} the 'mediation effect' and $E(\theta_{NIDr})$ the 'average mediation effect'.

In contrast to specifications based on potential outcome variables, the above causal effects have been presented in terms of conditional distributions under a decision theoretic framework, where the mediator variable is manipulated by a non-random intervention variable irrespective of the actual assignment of the randomised intervention variable.[35] This intervention variable for the mediator accommodates the two types of manipulation that distinguish the above controlled and natural effects: the intervention variable for the mediator either fixes the mediator variable at a specific level or specifies a distribution from which the mediator is drawn. For the manipulation based on specifying a distribution for the mediator, two types of distributions were considered by Geneletti:[35] (1) a distribution conditional on a fixed level of the randomised intervention (not necessarily the assigned level) and (2) a distribution that is not indexed by the intervention but instead a distribution that is 'appropriate for the context' such as flipping a coin to decide on the value for the mediator. The mediator distribution indexed at level of the randomised intervention yields the above natural effects, θ_{NDr} and θ_{NIDr}, whereas the distribution not indexed at a level of the intervention yields a different type of natural direct effect. While the natural indirect effect is identified if the natural direct effect is identified under this non-treatment indexed distribution, it is not clear what the indirect effect means given the manipulation is not performed in terms the set level of the randomised intervention as it is in (4) above.

The PS approach offers a different definition of the direct effect an alternative to the controlled and natural effects, especially when the mediator is not manipulable. With a binary mediator (e.g. non-study therapy), four possible principal strata exist and have been interpreted as follows:[30–32] for the first principal stratum $C=1$, the participant would exhibit a positive level for the mediator (e.g. no non-study therapy) if the patient were to be randomly assigned to the intervention arm (e.g. CBT), and *vice versa* if the patient were to be assigned to the comparison arm (e.g. usual care). That is, $M_r=r$ for $r=0, 1$. For the second principal stratum $C=2$, the participant would exhibit a negative level for the mediator (e.g. non-study therapy) if the participant were to be randomly assigned to the intervention arm, and *vice versa* if the patient were to be assigned to the comparison arm (i.e. $M_r=1-r$ for $r=1, 0$). For the third principal stratum ($C=3$), the participant would exhibit a negative level for the mediator regardless of randomisation status (i.e. $M_r=0$ for $r=1, 0$). For the fourth principal stratum ($C=4$, the

participant would exhibit a positive level for the mediator regardless of randomisation status (i.e. $M_r = 1$ for $r = 1, 0$). In the two principal strata for which the prospective mediator factor behaviour is held constant when changing intervention conditions (i.e., $M_1 = M_0 = m$ for $m = 0, 1$), the randomised intervention effects within these 'fixed mediator' principal strata (third and fourth strata above) are controlled direct effects but in sub-populations not the total population, unlike $\theta_{CD\ m}$.[30,31]

Given an observed baseline assignment status and also observed mediator level, each participant can potentially belong to either of two of the four principal strata, which defines the likelihood as a mixture likelihood. For example, a participant randomised to CBT and who did not seek non-study therapy would belong to either the principal stratum that never seeks non-study therapy or the one that does not seek non-study therapy only under CBT. In contrast, a participant randomised to the comparison group and who did not seek non-study therapy would belong to either the principal stratum that never seeks non-study therapy or the one that seeks non-study therapy only under CBT.

Under this principal stratification context, direct effects have been proposed and indirect effects have been discussed.[30–32] where $c = 1 \ldots 4$ for the four principal strata. Such effects are specified in terms of the ITT effects within each principal stratum: $E[Y_1 - Y_0|C = c]$ such that the weighted average of these effects equals the overall ITT effect, where the weights are the probability of membership in a principal stratum, π_c: $E[Y|R = 1] - E[Y|R = 0] = \sum_c E[Y_1 - Y_0|C = c]\pi_c$. The ITT effects within the principal strata for which the participants do not change their mediator behaviour (i.e. $C = 3, 4$ where $M_1 = M_0 = m$ for $m = 0, 1$) are direct effects of the randomised intervention but only for those principal strata, but not the other strata:

$$E[Y_1 - Y_0|C = 3] = E[Y_1 - Y_0|M_1 = M_0 = 0] \tag{5}$$

$$E[Y_1 - Y_0|C = 4] = E[Y_1 - Y_1|M_1 = M_0 = 1]. \tag{6}$$

In relating the controlled and natural effects to the above PS effects, VanderWeele[31] formally showed that the absence of controlled or natural direct effects is a stronger condition than no PS direct effects. Finally, it is not possible to define causal indirect effects in the presence of unmeasured confounding of differences across the principal strata. This is because isolating indirect effects requires contrasts among principal strata rather than within principal strata.

4 Non-parametric identifiability and estimation of average causal effects

All of the individual-level causal effects in (1)–(4) involve contrasts between potential outcome variables for an individual and correspond to average contrasts between potential outcome variables or average causal effects: $E(\theta_{CD\ m})$, $E(\theta_{ND})$, $E(\theta_{CM\ r})$ and $E(\theta_{NID})$, respectively. Robins[5] and VanderWeele and Vansteelandt[7] noted that $E(\theta_{NID0}) = E(\theta_{ITT}) - E(\theta_{ND1})$, where $E(\theta_{ITT}) = E(Y_{1\,M_1}) - E(Y_{0\,M_0})$ is the 'total' or ITT effect. Hence, any effect of the treatment on outcome not totally direct is an indirect contribution. Relationships between the average natural causal effects and the ITT effect can be obtained under the composition assumption.[7] It states that the potential outcome Y_r with the randomised intervention set to r equals the potential outcome $Y_{r\,M_r}$ with the randomised intervention set to r and the potential mediation variable M_r set to the value it would have been if R had been set to r.

Non-parametric identifiability of these average causal effects is achieved when we can specify them as functions of observed data distributions under certain assumptions. We review different

combinations of assumptions that lead to such identifiability before presenting these average causal effects in terms of observed data.

4.1 Assumptions

The identifiability assumptions resolve two impossibilities. First, there is the impossibility of observing both potential outcomes of a causal contrast for an individual (e.g. Y_{1m} and Y_{0m}). Identifiability of average controlled and natural causal effects requires resolution of this impossibility with assumptions. Second, it is impossible to observe the individual potential outcome variable simultaneously corresponding to two different levels of R, $Y_{rM_{r'}}$, where $r \neq r'$ for all m. Only the average natural effects, $E(\theta_{NDr})$ and $E(\theta_{NIDr})$, require assumptions to resolve this impossibility. The discussion below distinguishes the assumptions necessary for resolving these two impossibilities, and in turn the assumptions necessary for identifying the controlled and natural causal effects. In addressing the second impossibility (two different levels of the intervention for a single potential outcome), the literature focuses on identifying the natural effects with different variations of ignorability for both the baseline intervention and the mediator and also assumptions of no-interaction between them. We review these different combinations of assumptions below for the natural effects while also presenting the subset of assumptions that are needed for identifying the controlled effects. However, before reviewing these assumptions, we first review the more basic assumptions needed for non-parametric and parametric estimation of causal effects. After reviewing the assumptions, we then review non-parametric estimation.

Identifiability and estimation of the average causal effects from observed data first require two assumptions under what is sometimes called the stable unit treatment value assumptions (SUTVA). The 'no interference between study units' part of SUTVA is needed to use the above potential variable notation with scalar indices rather than vector indices representing randomised intervention assignment and/or mediation status of other subjects. That is, Y_{rm} is used rather than $Y_{\mathbf{r}\ \mathbf{m}}$ where $\mathbf{r}$ and $\mathbf{m}$ are the vectors of manipulated randomised intervention and mediator levels for all subjects. Departures from this assumption may occur when interventions such as behavioural or educational interventions are administered at the primary practice or provider level, such as in our examples. For example, when a provider administers the intervention to encourage depressed patients to take prescribed treatment for depression, the provider may learn from previous study patients and apply what he or she learns to subsequent study patients.

The consistency assumption of SUTVA is needed for estimation by linking the potential outcomes to the observed outcomes. In words, the consistency assumption implies that the observed random variable will equal one of the corresponding potential random variables even if the administration of the randomised intervention and mediation behaviour vary slightly.[36] In the case of a binary intervention and mediator, we can write under the consistency assumption: $Y = r\,m\ Y_{rm} + (1-r)\ m\ Y_{1-r\,m} + r\ (1-m)\ Y_{r\ 1-m} + (1-r)\ (1-m)\ Y_{1-r\ 1-m}$ for $r = 0$, 1 and all m. The consistency assumption is violated when there are different versions of a randomised intervention not reflected in the variable notation. Such violations may occur when there are different forms of administration such as interactions between the provider and patients through phone or in-person contact.

Under SUTVA, a variety of different combinations of ignorability and interaction assumptions have been proposed to achieve such non-parametric identifiability of the controlled and natural causal effects. Good reviews were provided by Imai et al.[8] and van der Laan and Petersen.[1] We first consider the assumptions under the absence of post-baseline confounders, addressing identifiability under post-baseline confounders in Section 6. The assumptions for identifying the controlled direct

effects are well established, whereas there are different sets of assumptions in the literature for identifying the natural direct and indirect effects. We review such assumptions from Pearl,[4] Robins,[5] van der Laan and Petersen,[1] Hafeman and VanderWeele,[37] and Imai et al.[8] Instead of presenting these groups of assumptions separately, we present them in terms of three components: (1) ignorability of the randomised intervention; (2) ignorability of the mediator; and (3) additional assumptions such as no-interaction between the randomised intervention and mediator. Finally, we note that it is not possible to use observed data to assess or distinguish between the different sets of identifiability assumptions; therefore, decisions regarding which set should be used should be based on underlying scientific issues.

To better illustrate the assumptions necessary for non-parametric identifiability of the average causal effects, we specify them as functions of the following observable quantities without specifying a parametric model: (1) the conditional means for the outcome, $E(Y \mid R=r, M=m, \mathbf{X}=\mathbf{x})$ for both the controlled and natural effects; (2) the conditional distribution of the mediator, $F_{M|R=r,\mathbf{X}=\mathbf{x}}(m)$, for the natural effects; and the distribution of the baseline covariates, $F_X(x)$, for all effects:

$$E(\theta_{CDm}) = \int [E(Y|R=1, M=m, \mathbf{X}=\mathbf{x}) - E(Y|R=0, M=m, \mathbf{X}=\mathbf{x}]\, \mathrm{d}F_{\mathbf{X}}(x) \tag{7}$$

$$E(\theta_{CMr}) = \int [E(Y|R=r, M=m, \mathbf{X}=\mathbf{x}) - E(Y|R=r, M=m', \mathbf{X}=\mathbf{x}]\, \mathrm{d}F_{\mathbf{X}}(x) \tag{8}$$

$$E(\theta_{NDr}) = \int\int [E(Y|R=1, M=m, \mathbf{X}=\mathbf{x}) - E(Y|R=0, M=m, \mathbf{X}=\mathbf{x})]\, \mathrm{d}F_{M|R=0,\mathbf{X}=\mathbf{x}}(m)\, \mathrm{d}F_{\mathbf{X}}(x) \tag{9}$$

$$E(\theta_{NIDr}) = \int\int E(Y|R=0, M=m, \mathbf{X}=\mathbf{x})\big[\mathrm{d}F_{M|R=1,\mathbf{X}=\mathbf{x}}(m) - \mathrm{d}F_{M|R=0,\mathbf{X}=\mathbf{x}}(m)\big]\mathrm{d}F_{\mathbf{X}}(x). \tag{10}$$

We see that the average natural direct effect of the randomised intervention is an average of the expectation contrast between randomised intervention levels with respect to the distribution of the mediator. In contrast, the average natural indirect effect of the randomised intervention is an average of the expectation contrast but between the mediator distributions for the different randomised intervention levels again averaged with respect to the distribution of the mediator. We also note that in the above equations, the ignorability of the randomised intervention (R) is conditional on $\mathbf{X}$. In the presence of post-baseline confounders, we condition the mean of Y and distribution of M on $\mathbf{Z}$: $E(Y \mid R=r, M=a, \mathbf{Z}=\mathbf{z}, \mathbf{X}=\mathbf{x})$ and $F_{M|R=r,\mathbf{Z}=\mathbf{z},\ \mathbf{X}=\mathbf{x}}(m)$, and then integrate over the distribution of $\mathbf{Z}$ with respect to its conditional distribution: $F_{\mathbf{Z}|R=r,\ \mathbf{X}=\mathbf{x}}(z)$.

The following sequential ignorability results pertain primarily to the identifiability of the natural effects in (9) and (10), whereas the identification of the controlled effects in (7) and (8) uses less restrictive assumptions since the potential outcome variable corresponding to both levels of R, $Y_{rM_{r'}}$ is not involved. In particular, identification of $E(\theta_{CDm})$ and $E(\theta_{CMr})$ requires just (14) and (19), which do not involve sequential ignorability assumptions involving both r and r'.

The first aspect of sequential ignorability pertains to the ignorability of the randomised intervention with respect to both the mediator and outcome. Pearl[4] defined this to be:

$$E(Y_{rm}|\mathbf{X}=\mathbf{x}) \quad \text{and} \quad E(M_{r'}|\mathbf{X}=\mathbf{x}) \quad \text{are identifiable} \tag{11}$$

for r, $r'=0$, 1. To make the identifying assumptions more comparable among different authors, we assume (11) and the assumptions below apply to both levels r or all combinations of r and r'

when relevant. While the randomised intervention variable is not explicitly specified in the above definition, it is specified explicitly such as that specified by Imai et al.:[8]

$$Y_{r'm}, M_r \perp R|\mathbf{X} = \mathbf{x} \tag{12}$$

for r, $r' = 0$, 1, and m. Note that the randomised intervention index differs for the potential outcome (r) and mediator variables (r'). Hence, the full joint distribution of all the potential outcome and mediator variables is specified. A less restricted assumption was made by van der Laan and Petersen[1] and Hafeman and VanderWeele:[37]

$$Y_{rm}, M_r \perp R \mid \mathbf{X} = \mathbf{x} \tag{13}$$

for $r = 0$, 1 and all m. An alternative and even less restrictive assumption was made by van der Laan and Petersen[1] with respect to the marginal distributions of the potential outcome and mediator variables:

$$Y_{rm} \perp R|\mathbf{X} = \mathbf{x} \tag{14}$$

$$M_r \perp R|\mathbf{X} = \mathbf{x} \tag{15}$$

for $r = 0$, 1 and all m.

The distinction between the assumptions necessary for identifying the controlled and natural effects starts with distinctions among the above assumptions involving the ignorability of the randomised intervention. The least restrictive assumption above for ignorability of the randomised intervention is (14). Depending on the pairing of one of the above assumptions with one of the versions of the ignorability assumption of the mediator described below, the controlled or natural effects may still not be identifiable. In such cases, other non-ignorability assumptions are needed.

The second aspect of sequential ignorability entails the ignorability of the mediator with respect to the outcome. Under the ignorability assumption for the randomised intervention in (11), Pearl[4] made the following ignorability assumption with the mediator for non-parametric identifiability of the average causal effects in (7)–(10), especially the natural causal effects:

$$Y_{rm} \perp M_{r'}|\mathbf{X} = \mathbf{x}. \tag{16}$$

for r, $r' = 0$, 1 and all m. Robins[5] referred to the non-parametric identifiability conditions (11) and (16) as the NPSE model. The key condition that leads to identifiability of the natural causal effects in (9) and (10) under (11) and (16) is the independence between the two potential variables corresponding to different set levels of the randomised intervention, r and r'. Given the impossibility of observing potential variables defined in terms two different levels of the randomised intervention (i.e., $Y_{r,Mr'}$), the above assumptions allow two substitutions: (1) substitution of a conditional expectation of Y_{rm} given $R = r$ and $M = m$ with the analogous conditional expectation of $Y_{r,Mr'}$ and (2) replacement of the conditional expectation of Y_{rm} with the conditional expectation of Y, leading to the Equations (7)–(10).

Hafeman and VanderWeele[37] specified a weaker second component of the sequential ignorability assumption than Pearl's assumption in (16) but still involving different indices (r and r') for R:

$$Y_{rm} \perp M | R = r', \mathbf{X} = \mathbf{x}. \tag{17}$$

With (12), (17) is sufficient for non-parametric identifiability of the average causal effects in (7)–(10), particularly the natural effects in (9) and (10).

As an alternative to the identifiability assumptions in (11) and (16) by Pearl[4] and (12) and (17) by Imai et al.,[8] van der Laan and Petersen[1] identified the average natural direct effect in (9) with assumptions (14), (15), (17), and an another identifying assumption:

$$E(Y_{1m} - Y_{0m} | M_0 = m, \mathbf{X} = \mathbf{x}) = E(Y_{1m} - Y_{0m} | \mathbf{X} = \mathbf{x}) \tag{18}$$

van der Laan and Petersen[1] suggested that assumption in (18) is more plausible than the sequential ignorability assumption in (16) by Pearl.[4] For the latter assumption, the individual potential outcome, Y_{1m} is independent of the potential mediator outcome at the reference randomised intervention level, M_0, as opposed to the former assumption that the causal difference $Y_{1m} - Y_{0m}$ is mean independent of M_0. Regardless, this latter assumption offers an introduction to identifying assumptions based on interactions, given it is a type of no-interaction assumption in that the average controlled direct effect is constant for all levels of the potential mediator variable under assignment to level 0 of the intervention. We now consider other no-interaction assumptions for identifiability.

Using an individual-level no-interaction assumption, Robins[5] achieved non-parametric identifiability of the average natural causal effects in (9) and (10) by assuming Pearl's (11) first component of the sequential ignorability assumption and then relaxing (17) such that the ignorability of M only occurs for the same level of the randomised intervention, $r = r'$:

$$Y_{rm} \perp M | R = r, \mathbf{X} = \mathbf{x}. \tag{19}$$

However, because of the absence of an assumption about different levels of r and r' for the natural effects, Robins and Greenland[34] and Robins[5] proposed the following individual-level no-interaction assumption under which the controlled intervention direct effect, θ_{CDm}, to be constant for all m for identifiability:

$$Y_{rm} - Y_{r'm} = Y_{rm'} - Y_{r'm'} = B(r, r') \tag{20}$$

for $r = 1$, $r' = 0$ and all $m \neq m'$. Here $B(r, r')$ is independent of m, indicating no interaction at the individual level. Robins[5] referred to the sequential ignorability assumptions (11) and (19) as the FRCISTG model.

We note that (19) is the least restrictive conditional ignorability assumption among those considered above for the mediator, as (13) is for the randomised intervention. These two assumptions are sufficient for identifying the controlled direct effects of the randomised intervention and mediator, θ_{CDm} and θ_{CMr} for given m and r, respectively. Finally, the no-interaction assumption in (20) also implies that the controlled intervention direct (θ_{CDm}) and mediation (θ_{CMr}) effects are constant across m and r, respectively.

As an alternative to the no-interaction assumption in (20) involving contrasts between just potential outcomes, Hafeman and VanderWeele[37] (Equation A3) presented another type of

identifying no-interaction assumption for binary mediators in addition to the assumptions in (13) and (17):

$$E(Y_{1m} - Y_{0m}|R = 1, M = 1, \mathbf{X} = \mathbf{x}) = E(Y_{1m} - Y_{0m}|R = 1, M = 0, \mathbf{X} = \mathbf{x}) \quad (21)$$

for $m = 0, 1$. A reviewer suggests that this assumption may constrain observed data in such away that they may be partially testable.

In Section 7.1, (21) and a number of similar no-interaction assumptions are made under the rank preserving model (RPM) without assuming ignorability of the mediator. As (21) illustrates, this set of no-interaction assumptions specifies that either the average controlled direct effect of the randomised intervention or the mediator conditional on the observed randomised intervention (R) and mediator variables (M) does not differ across different groups of subjects distinguished by their observed randomised intervention and mediation levels. For example, (21) constrains the average controlled direct effect of the randomised intervention to be constant across subject groups differentiated by the levels of the binary mediator. In Section 7.1 under the RPM, this assumption applies to continuous mediators.

With the assumption in (13), (17) and (21), Hafeman and VanderWeele[37] still required an additional assumption to identify the natural effects (A2 in Hafeman and VanderWeele):[37]

$$E(Y_{1m}|R = 1, M = m, \mathbf{X} = \mathbf{x}) = E(Y_{1m}|R = 0, M = m, \mathbf{X} = \mathbf{x}) \quad (22)$$

for all m.

Under the decision theoretic framework with conditional distributions rather than potential outcomes, Geneletti[35] made what is equivalent to the least constrained sequential randomisation assumptions in (13) and (19) augmented by a conditional independence assumption between the randomised intervention and mediator given the intervention variable that manipulates the mediator. This assumption is analogous to the above assumptions that resolve the impossibility of observing the potential outcome, $Y_{rM_{r'}}$, given it involves two different levels of the randomised intervention.

4.2 Non-parametric estimation

We now review estimation of the controlled and natural effects in (7)–(10) under the corresponding assumptions for identifiability.[4,5,8,34,37] First, $E(Y \mid R = r, M = a, \mathbf{X} = \mathbf{x})$ in (7)–(10) is replaced with its corresponding sample-based average of Y for the group of participants with $R = r$, $M = m$ and $\mathbf{X} = \mathbf{x}$, $\hat{\mathrm{E}}(Y \mid R = r, M = a, \mathbf{X} = \mathbf{x})$, for both the controlled and natural effects. In addition, for the conditional distributions of the mediator and baseline covariates, we define the indicator variable $I(expression) = 1$ when expression is true and $I(expression) = 0$ otherwise. We estimate the conditional distribution of the mediator, $F_{M|R=r,\mathbf{X}=\mathbf{x}}(m)$, with the empirical distribution of M for the sample of subjects with $R = r$ and $\mathbf{X} = \mathbf{x}$: $\hat{f}_{M|R=r,\mathbf{X}=\mathbf{x}}(m) = \sum \frac{I(M=m,\ R=r,\ \mathbf{X}=\mathbf{x})}{I(R=r,\ \mathbf{X}=\mathbf{x})}$. We also estimate the conditional distribution for $\mathbf{X}$, $F_{\mathbf{X}|R=r}(\mathbf{x})$ with the empirical distribution of $\mathbf{X}$ given $R = r$, $\hat{f}_{\mathbf{X}|R=r}(\mathbf{x}) = \sum \frac{I(\mathbf{X}=\mathbf{x},\ R=r)}{I(R=r)}$, for all effects:

$$\hat{E}(\theta_{CDm}) = \sum_{\mathbf{x}} \left\{ \hat{E}(Y|R = 1, M = m) - \hat{E}(Y|R = 0, M = m) \right\} \hat{f}_{\mathbf{X}}(\mathbf{x}) \quad (23)$$

$$\hat{E}(\theta_{CMr}) = \sum_{\mathbf{x}} \left\{ \hat{E}(Y|R = r, M = m) - \hat{E}(Y|R = r, M = m') \right\} \hat{f}_{\mathbf{X}}(\mathbf{x}) \quad (24)$$

$$\hat{E}(\theta_{NDr}) = \sum_{\mathbf{x}} \sum_{m} \left\{ \hat{E}(Y|R=1, M=m) - \hat{E}(Y|R=0, M=m) \right\} \hat{f}_{M|R=r,\mathbf{X}=\mathbf{x}}(m) \hat{f}_{\mathbf{X}}(\mathbf{x}) \tag{25}$$

$$\hat{E}(\theta_{NIDr}) = \sum_{\mathbf{x}} \sum_{m} \hat{E}(Y|R=r, M=m) \left\{ \hat{f}_{M|R=1,\mathbf{X}=\mathbf{x}}(m) - \hat{f}_{M|R=0,\mathbf{X}=\mathbf{x}}(m) \right\} \hat{f}_{\mathbf{X}}(\mathbf{x}) \tag{26}$$

for $r=0$, 1 and all m. The more continuous M is and the higher the dimension of $\mathbf{X}$, the more variable are the empirical distributions. Consequently, specifying M to be binary and $\mathbf{X}$ to consist of one or two categorical variables will result in more stable empirical distributions, although limiting the dimension of $\mathbf{X}$ may lead to bias due to confounding.

5 Causal inference under the LSEM

In addition to identifying non-parametrically the average causal effects in (7)–(10), the above sets of assumptions also facilitate the use of the linear regression for continuous outcomes and mediators, sometimes referred to as LSEM but only when there is no post-baseline confounder, $\mathbf{Z}$. We address the LSEM in the presence of post-baseline covariates in Section 6. In the case without post-baseline covariates, Baron and Kenny[6] are frequently cited for the LSEM approach in the social science literature, as they presented a set of conditions that parallel the conditions for no-confounding in an observational study situation. Imai et al.[8] justified using this approach under the sequential ignorability assumptions in (12) and (17) and linearity of the models for the outcome and mediator, which are reviewed next.

5.1 Model and assumptions

The LSEM is presented below in terms of two linear regression models: one for the outcome (Y) with main effects for the randomised intervention (R); and one for the mediator (M) and for the mediator with a main effect for the randomised intervention main effect:

$$M = \alpha_0 + \alpha_1 R + \boldsymbol{\alpha}_2 \mathbf{X} + \epsilon_1 \tag{27}$$

$$Y = \beta_0 + \beta_1 R + \beta_2 M + \boldsymbol{\beta}_4 \mathbf{X} + \epsilon_2. \tag{28}$$

Identifiability of the above parameters requires the sequential ignorability assumptions for the controlled effects in (7) and (8), which are (14) and (19). However, the above LSEM makes the additional individual-level no-interaction assumption by Robins,[5] which Imai et al.[8] noted can be relaxed as they show with an LSEM including an interaction term between the randomised intervention and mediator.

Under (14) and (19) and the above LSEM with no individual-level interaction, the controlled and natural direct effects of the randomised intervention and the controlled effect of the mediator are specified in terms of the parameters in (28) as:

$$\begin{aligned} E(\theta_{CDm}) &= E(\theta_{NDr}) \\ &= \beta_1 \\ E(\theta_{CMr}) &= \beta_2 \end{aligned} \tag{29}$$

for $r=0$, 1 and all m.

Under Imai's more restrictive sequential ignorability assumptions (12) and (17), the average natural indirect effect in (10) equals the following parameterisations under the LSEM:

$$E(\theta_{NIDr}) = \beta_2\alpha_1 \tag{30}$$

for $r = 0, 1$. The above equality also holds under the other sets of sequential ignorability and/or no-interaction assumptions by Pearl[4] and Robins.[5] If the mediator, M, is binary, then the model for M in (27) needs to take this into account such as with a logit link. However, this precludes the specification of the product of parameters for the average causal indirect effect in (30). The alternative estimation approach is the non-parametric estimator of the natural indirect effect in (26). Hafeman[38] showed that the proportion explained calculated using a logit link is a biased weighted average of the natural indirect effects.

5.2 Estimation under the LSEM

Estimation of the average causal effects under the LSEM depends on the absence of post-baseline confounders, **Z**. Without such confounders under the assumptions discussed in Section 5.1, estimation and inference for the average controlled and natural effects of the baseline intervention in (7) and (9) and for the average controlled effect of the mediator on outcome in (8) proceed with ordinary least squares (OLS) regression. Estimation and inference for the average natural indirect effect in (10) can be basedon method of moments estimation and the delta method.[14] More specifically, under the individual-level no-interaction LSEM in (28) and (28) with sequential ignorability assumptions (14) and (19), the average natural indirect effect can be estimated as:

$$\hat{E}(\theta_{NIDr}) = \hat{\beta}_2\hat{\alpha}_1 \tag{31}$$

for $r = 0, 1$. MacKinnon et al.[4] reviewed a number of two different general classes of approaches for testing mediation: (1) Wald tests that the indirect effect equals zero $E(\theta_{NIDr}) = 0$ based on the above estimator and different standard errors, and/or test distribution estimators; (2) Wald tests that the averaged controlled direct effect equals zero: $E(\theta_{CDm}) = E(\theta_{NDr}) = 0$. The null and alternative hypotheses of these two approaches relate in opposite ways to the presence or absence of mediation (i.e. null applies to mediation in the latter case but to the presence of mediation in the former case, and conversely for the alternative hypotheses). Nonetheless, MacKinnon et al.[14] showed that these different tests vary in power and test size (maximum probability of type I error) under the no-interaction LSEM in (27) and (28).

6 LSEM with post-baseline confounders

When the randomised intervention, R, impacts the post-randomisation confounders, **Z**, average natural effects are not identified and a G-computation or sequential G-estimation approach is needed to estimate the average controlled direct effect of R.[2,7,39] Such covariates are included only in the outcome component of the LSEM in (28). Under this extended LSEM with the randomised intervention identifiability assumption (12) and an extension of the mediator identifiability assumption (19) to condition on **Z**, the OLS regression estimator of the average controlled effect of the mediator is not biased, whereas the OLS regression estimator of the average controlled effect of the randomised intervention is biased. In this case, the mediator is

impacted by both the randomised intervention and the post-baseline factor, both of which also impact directly the outcome. In contrast, one can obtain asymptotically unbiased estimators of the average controlled effect of the mediator based on OLS regression methods under the LSEM. This is possible, because the post-baseline variables, $\mathbf{Z}$, are confounders with respect to the effect of the mediator on outcome, and therefore OLS regression methods are appropriate under the no-interaction LSEM with sequential ignorability assumptions, conditioning on the post-baseline confounder. Finally, when R does not impact $\mathbf{Z}$, the OLS estimator of the average controlled direct effect of R is not biased.

When R impacts $\mathbf{Z}$, Vansteelandt[2] and Joffe and Greene[40] have resolved the bias issue for the average controlled direct effect of the randomised intervention by using a two-stage OLS procedure. We review this approach, but first presenting the model and assumptions for identifiability and unbiased estimation of the model parameters as causal parameters corresponding to the controlled direct effects in (7) and (8).

6.1 Model and assumptions

The LSEM with post-baseline confounders is a straightforward extension of the LSEM with no post-baseline confounders in (27) and (28) by inclusion of main effects parameters for the post-baseline confounders, $\mathbf{Z}$, in only the outcome model in (28):

$$M = \alpha_0 + \alpha_1 R + \boldsymbol{\alpha}_2 \mathbf{X} + \epsilon_1 \tag{32}$$

$$Y = \beta_0 + \beta_1 R + \beta_2 M + \boldsymbol{\beta}_3 \mathbf{X} + \beta_4 \mathbf{Z} + \epsilon_2. \tag{33}$$

The randomised intervention ignorability component of the sequential ignorability assumption in (14) is the same for the no-interaction LSEMs with or without post-baseline confounders, although the mediation component of the sequential ignorability assumption is almost identical to (19) under the no-post-baseline confounder LSEM except for conditioning on $\mathbf{Z}$:

$$Y_{rm} \perp M | \mathbf{X} = \mathbf{x}, R = r, \mathbf{Z} = \mathbf{z} \tag{34}$$

for $r = 0$, 1 and all m. Under a specified LSEM with post-baseline confounders, sequential ignorability assumptions (14) and (34) identify the average controlled causal effects (7) and (8) in terms of the model parameters in (32) and (33).

6.2 Estimation under the LSEM

Under the assumptions in (14) and (34), $E(\theta_{CMr}) = \beta_2$, of which the unbiased OLS estimator is used in two-stage estimation of β_1, the average controlled direct effect of the intervention under the second stage model in (33).[2] That is, the estimator of $\beta_2 = E(\theta_{CMr})$, say $\hat{\beta}_2$, is used to adjust Y for M in estimating the direct effect of R: $Y_{adj} = Y - \hat{\beta}_2 M$. This adjusted Y is then used as the dependent variable in an OLS regression on the randomised intervention:

$$Y_{adj} = \gamma_0 + \gamma_1 R. \tag{35}$$

The resulting OLS estimator of γ_1 is an unbiased estimator of the average controlled direct effect of the intervention, $E(\theta_{CDm}) = E(\theta_{NDr}) = \beta_1$ for all m.

Vansteelandt[2] also extended this approach to include an interaction between the randomised intervention and mediator. Only estimation and testing of the main effect of the randomised intervention requires the following adjustment based on a revisited Y_{adj}, $Y_{adj} = Y - \hat{\beta}_2 M - \hat{\beta}_4 RM$, where β_4 is the interaction parameter. Estimators of the causal parameters under the above approach may still be biased if R and M are not independent. Note that in contrast, mediation requires that R have an effect on M. The amount of bias under associations between the randomised intervention and mediator is a subject of current research.[41] Kraemer et al.[16] make this point but provide no analytical or simulation evidence to support it.

Vansteelandt[2] offered an alternative doubly–robust estimation approach as well as an excellent review of methods for estimating the average causal effects in the context of a post-baseline confounder. The doubly robust 'sequential G-estimation' approach, which requires specification of inverse probability weights and multiple components for the estimating equation, is doubly robust under mis-specification of either the mediator or outcome model. In terms of the LSEM, this would mean mis-specification of either the mediator model in (27) or the outcome model in (28). However, because of instability due to the weights and also components of the estimating function, they recommend an unweighted G-estimation approach, which was also used by Joffe and Greene.[40] Other methods include G-computation based on contrasts of averages of conditional expectations from standard regression with respect to the observed distribution of the post-baseline confounder. However, this approach can be cumbersome with several continuous post-baseline confounders. In addition, it does not yield model parameters corresponding to the average causal effects in (7)–(10). Alternatively, marginal structural models estimated with inverse probability weighting have been implemented to adjust for post-baseline confounding in the mediation context. However, such an estimation approach is vulnerable to large weights arising from either few subjects for specific mediator levels as in the case of continuous mediators or strong relationships of the mediator with confounders and/or the randomised intervention. Consequently, Joffe and Greene[40] and Vansteelandt[2] recommenced the above two-stage procedure.

7 Causal mediation without sequential ignorability

We now present several causal mediation approaches without any assumptions about the ignorability of the mediator, but still assuming the ignorability of the randomised intervention. The ignorability of the mediator assumptions is replaced by other identifying assumptions discussed below. The causal approaches include extensions of the semi-parametric approach based on G-estimation[12] and likelihood or Bayesian-based PS[29] approaches. The average controlled direct effects of the randomised intervention and mediator are estimated under the semi-parametric approach. The PS yields direct effects of the randomised intervention in certain principal strata or latent classes based on potential mediation behaviour under each randomised intervention level. We present each of these two causal mediation strategies and their respective assumptions, separately, given that they are very different approaches requiring different types of assumptions.

7.1 Semi-parametric approach

The semi-parametric approach for continuous outcomes with G-estimation has been presented in the causal literature as the SNMM in the context of estimating the causal effects of adherence in randomised treatment trials.[42] Such approaches typically make an assumption about no direct effect of the randomised intervention on outcome (i.e. $\theta_{CDm} = 0$) under the exclusion restriction.

We present a similar approach for the mediation context but of course without the exclusion restriction. The mediation approach is based on the RPM used in Ten Have et al.[12] While it is not possible to use observed data to distinguish between the SNMM and RPM, the RPM makes additional assumptions regarding causal error structure that are not made by the SNMM approach. We take the RPM approach because it closely parallels in the LSEM in model format and thus makes the two approaches easier to compare.

7.1.1 Assumptions for semi-parametric approach

Under this approach to estimating the average controlled direct effects of the randomised intervention and mediator $E(\theta_{CDm})$ and $E(\theta_{CMr})$, we make the following ignorability assumption for the randomised intervention, which is satisfied in randomised trials:

$$E(Y_{r'm}|R = r, \mathbf{X} = \mathbf{x}) = E(Y_{r'm}|\mathbf{X} = \mathbf{x}). \tag{36}$$

for r, $r' = 0$, 1. This assumption is almost equivalent toassumptions (1) and (5) in Hernan and Robins,[27] except we condition on $\mathbf{X}$. For the no-confounding assumption of the randomised intervention, (36) for the RPM is a less restrictive assumption than the LSEM assumption in (15) because of the absence of assumptions involving the potential mediator variable, M_r, in (36). Nor does the RPM require sequential ignorability as represented by the no-confounding assumption in (15), which is required by the LSEM either with or without the post-baseline confounder, $\mathbf{Z}$.

However without sequential ignorability, certain no-interaction assumptions are needed for the RPM. These take two forms. The first is the no-structural or causal interaction imposed by Robins[5] in (20) above. This assumption says that either the controlled causal effect of the randomised intervention (θ_{CDm}) does not differ across different 'set' levels (m) of the mediator; or that the controlled effect of the mediator (θ_{CMr}) does not differ across different 'set' levels (r) of the randomised intervention. These structural interactions can be tested using the G-estimation approach described below for the semi-parametric mediation model, as indicated in Joffe et al.[43]

The second set of no-interaction assumptions represent an extension of the no-interaction assumptions in Equations (6a), (6b), (8a) and (8b) of Hernan and Robins[27] in the IV adherence context under the exclusion restriction (no direct effect of the IV (e.g., randomised intervention). This second set of no-interactions specifies that either the average controlled direct effects of the randomised intervention or the mediator conditional on the observed randomised intervention (R) and mediator variables (M) do not differ across different groups of subjects distinguished by their observed randomised intervention and mediation levels. The assumption in (21) represents one of these no-interaction assumptions although the mediator can be continuous under the rank preserving model. The testability of these no-interaction assumptions is a subject of future research. At best, they can be partially tested.

Finally, there are additional assumptions regarding the error and mean model structural under the RPM approach, which are described in Section 7.1.2. These assumptions along with both of the no-interaction assumptions in (20) and the extension of the Hernan and Robins[27] to the mediation context replace the various assumptions of ignorability of the mediator in (16)–(19).

Additionally, conditions for more precise estimation are presented in terms of interactions between the baseline covariates and the randomised intervention with respect to their impact on the mediation factor. The stronger the impact of the baseline covariates on the mediator, the more efficient the estimators of the causal parameters with the G-estimation under the RPM. We now describe the model and these conditions for more precise estimates.

7.1.2 Rank preserving model for semi-parametric approach

For the semi-parametric approach, the RPM is useful in that its structure is analogous to the structure of the LSEM in (28) above, but yet the RPM approach yields the same inference and estimation procedure as would the SNMM approach estimated in a similar manner.[12,44] By analogy with the LSEM in (28), we have the following RPM for all of the potential outcomes denoted by Y_{rm}. In the case of two levels for R and M, we need separate causal models for each of the four potential outcomes (Y_{11}, Y_{10}, Y_{01} and Y_{00}), in contrast to the standard linear regression model in (28), where we have only one model for Y:

$$Y_{rm} = g(\mathbf{x}) + \gamma_M m + \gamma_R r + \epsilon \tag{37}$$

for $r = 0$, 1 and all m and where $E(\epsilon \mid \mathbf{X}) = 0$. There are several points to note with respect to the RPM in (37). First, $g(\mathbf{X})$ is a function of the baseline covariates that does not need to be specified correctly for the estimator of the average direct intervention effect in (7) to be asymptotically unbiased. However, the degree of accuracy of the specification of $g(\mathbf{X})$ does impact the level of precision (efficiency) of this causal estimator.[26] Second, we note the constancy of error in (37) across the set indices r and m. This assumption does not impact the resulting estimators, as it has no relation to the observed data. However, it does allow us to define a linear model that is comparable in form to the LSEM in (28).

Under the above assumptions and model specifications for a specific individual participant, the two main effect parameters in the RPM, γ_R and γ_M are equal to the respective average controlled direct effects for the randomised intervention and mediator in (1) and (2):

$$\theta_{CDm} = \theta_{NDr} = \gamma_R \tag{38}$$

$$\theta_{CMr} = \gamma_M \tag{39}$$

for $r = 0$, 1 and all m. We now discuss in Section 7.1.3 G-estimation of the model parameters $\boldsymbol{\gamma}^T = (\gamma_R\ \gamma_M)$.

7.1.3 G-estimation procedure for semi-parametric approach

The weighted G-estimation procedure presented by Ten Have et al.[12] produced asymptotically unbiased estimators of γ_M and γ_R and corresponding standard errors under (37). That is, while these estimators may be somewhat biased for small sample sizes, this bias goes to zero as the sample size gets larger. Ten Have et al.[12] showed with simulations that this estimation procedure produces accurate inference under two separate samples sizes ranging from 100 to 300. Estimation is implemented using G-estimation equations.[26] The G-estimation equations represent extensions of randomisation tests, relying on the correctly specified distribution for the randomised assignment of the randomised intervention, but also requiring a mapping of the observed outcome Y to the potential outcome $Y_{0\ 0}$ by subtracting off the estimated linear combination of parameters and observed values of R, M and $\mathbf{X}$ in (37). A two-dimensional weight vector for each participant is incorporated into the G-estimation equations to obtain non-collinear identifying equations for each of the causal parameters, γ_R and γ_M. Because of the correct specification of the randomisation model, the resulting, identifying G-estimation estimating equations have zero expectation given R. Accordingly, they yield consistent estimators of $\boldsymbol{\gamma}$ without assuming randomisation of M but under the other assumptions described above. The variance–covariance for $\hat{\boldsymbol{\gamma}}$ is estimated after convergence of the G-estimation algorithm with a sandwich estimator described in Ten Have et al.[12]

The specification of the weight elements is crucial in two ways in terms of defining the separate estimating equations for each causal parameter and also in terms of efficiently of the estimators. To maintain identifiability under the above assumptions, it is imperative that collinearity between the elements of the weights is minimised. Specifically, each element of $\mathbf{W}(\tilde{\mathbf{x}})$ corresponds to a separate identifying equation for the corresponding structural parameter under certain assumptions. These weight elements can be derived from efficient score criteria in Robins et al.[45] under the linear structural distribution model given additional assumptions such as sequential ignorability and normal errors with constant variance. Under these assumptions, one of the weights requires strong baseline covariate modification of the randomised intervention effect on the mediator.

7.2 Principal stratification

As noted in Section 3, the PS approach relies on estimating the ITT intervention effect within latent sub-groups (i.e., principal strata) of participants who would naturally not change their mediator level regardless of the randomised intervention assignment (e.g. someone who would seek non-study psychotherapy regardless of whether they had CBT or not). This stratification occurs on the basis of their potential mediator behaviour under each of the two randomised intervention arms. Assuming the mediator is binary, two of the resulting four strata correspond to sub-classes of participants who would not change their mediator behaviour if their randomised intervention assignment changed. Hence, the mediator is controlled for participants in these two separate classes, and as a result, the estimated ITT intervention effect in each of these classes is the direct effect of the intervention at least for these sub-groups of participants.[30,31,46]

7.2.1 Assumptions for PS approach

PS models have often been identified, especially in the context of adherence to randomised treatment contexts, by a monotonicity assumption and then an exclusion restriction.[47] Under the monotonicity assumption, the principal stratum that is analogous to the defier principal stratum in the adherence context does not exist. Several forms of the exclusion restriction have been specified relating to the absence of a randomised intervention direct effect in principal strata in which the participants would not change their mediator level regardless of the presence or absence of the randomised intervention (i.e. the mediator is held constant in these participants).[48,49] That is, in this case, the exclusion restriction implies that the direct effect of the randomised intervention is zero in these principal strata in which participants do not change their mediator level. However, the exclusion restriction and monotonicity assumptions are not consistent with the goal of mediation analyses in that there is no reason to believe that any one of the principal strata does not exist (unlike in adherence contexts), and clearly mediation analyses would not be possible if direct effects of the randomised intervention were assumed to be absent in the 'fixed mediator' principal strata.

As a trade-off for monotonicity, the exclusion restriction and sequential ignorability, the PS approach for mediation requires assumptions involving strong covariate predictors of the principal strata and also parametric model assumptions for the outcome. Unlike the semi-parametric approach, the PS approach to mediation does not require the structural no-interaction assumption between R and M nor the no-interaction assumptions comparing different groups of subjects based on observed R and M, although it does require no interactions between the effect of R and baseline covariates $\mathbf{X}$ within principal strata. The separate direct effects of the randomised intervention in the principal strata for which the mediator level is held constant would be identified and estimated consistently within each principal strata when the covariates predict and distinguish

among these principal strata, but assuming constancy of the ITT effect of R across covariate levels within these two strata.

Also, unlike the RPM in (37), the error term in PS outcome model is assumed to have a fully parametric distribution, such as normal with mean zero and finite variance. Inference based on such models appears to be sensitive to these distribution assumptions.[48,50,51] Assuming a normal distribution for the outcome model error term, Imbens and Rubin[50] showed that the ITT effects in certain principal strata are biased under violations of this normality assumption. Separate variances may be assumed for the different principal strata under proper prior distributions to account for any departures from normality due to non-constant variance. Finally, because of the full likelihood or Bayesian approaches that is typically taken to fitting PS models, the ignorability of the randomised intervention is assumed to be in terms of the full probability distribution of the potential outcome and mediator variables, that is, the ignorability assumption in (13).

7.2.2 Model for the PS approach

Based on the principal stratum-specific ITT effects in defined in Section 3, the model that we consider for this mediation context is specified as follows for the potential outcome for the rth randomised intervention assignment and c-th principal stratum:

$$Y_r = \gamma_{PSc} r + \mathbf{x}^T \boldsymbol{\beta}_{PSc} + \epsilon_{rc}. \tag{40}$$

where $c = 1 \ldots 4$ for the four principal strata and for identifiability $\epsilon_{rc} \sim N(0, \sigma^2)$. Here $\boldsymbol{\beta}_{PSc}$ is the vector of covariate effects for the c-th principal stratum. The causal parameter γ_{PSc} is the ITT effect of the randomised intervention for the c-th principal stratum:

$$\gamma_{PSc} = E(Y_1|\mathbf{X} = \mathbf{x},\ C = c) - E(Y_0|\mathbf{X} = \mathbf{x},\ C = c).$$

The average direct effects of the randomised intervention correspond to the ITT effects of the randomised intervention (γ_{PSc}) in the 'fixed mediator' principal strata. Moreover, the standard ITT effect for the population equals the weighted sum of the stratum-specific ITT effects across all four strata with weights corresponding to probabilities of membership in each principal stratum, $\pi_c = \Pr(C = c|\mathbf{X} = \mathbf{x})$, such that $\sum_c \pi_c = 1$:

$$E[Y|\mathbf{X} = \mathbf{x},\ R = 1] - E[Y|\mathbf{X} = \mathbf{x},\ R = 0] = \sum_c E[Y_1 - Y_0|\mathbf{X} = \mathbf{x},\ C = c]\,\pi_c. \tag{41}$$

Because of the identifiability problems with relaxing the sequential ignorability, monotonicity and exclusion restriction assumptions, Bayesian techniques with informative priors have been used to fit PS models in the mediation context under the assumptions in Section 7.2.1 and the model in Section 7.2.2.[32] The relative informativeness of these priors depends on the predictive strength of the baseline covariates in the models for π_c.

7.2.3 Estimation under the PS approach

Estimation for the PS procedure is based on a mixture of distributions across principal strata. Specifically, each participant's likelihood is a mixture of two of the four densities corresponding to the possible principal strata given the observed M and R variables. Each observed outcome

density is a mixture of the potential outcome densities with weights as functions of probabilities of PS membership. The weights in the mixture densities are functions of the parameters of the prior multinomial distribution for the principal strata indicator variable C: $\eta_{c\mathbf{x}} = Pr(C = c \mid \mathbf{X} = \mathbf{x}) = \exp(\boldsymbol{\delta}_c^T \mathbf{x}) / \left[1 + \exp(\boldsymbol{\delta}_c^T \mathbf{x})\right]$, where $\boldsymbol{\delta}_c^T$ is a vector of parameters including an intercept parameter corresponding to the first element of $\mathbf{x}$, and $\boldsymbol{\delta}_1 = \mathbf{0}$ for identifiability. To obtain the posterior distributions of the model parameters, a MCMC algorithm may be implemented.[28,32]

8 Other mediation approaches

In addition to the above approaches for continuous outcomes with either binary or continuous mediators, methods have been proposed for non-continuous outcomes (e.g. binary, count or survival), sensitivity analyses under the LSEM, and more complex mediation contexts involving moderators (e.g. effect modifiers) or multiple mediators. Some work has been performed in all of these areas to different degrees mostly in the context of some form of sequential ignorability and/or no-interaction assumptions.

Much of the above results in Sections 4–6 for continuous outcomes apply to count and binary outcomes with some exceptions for binary outcomes under the logistic model. Specifically, in the absence of an intermediate confounder, $\mathbf{Z}$, affected by the randomised intervention, R, the non-parametric identifiability results for average controlled and natural additive effects in Section 4 apply to both count and binary outcomes.[2,52,53] Under loglinear or logistic model-based inference, the identifiability results in Section 5 with and without $\mathbf{Z}$ but assuming R does not impact $\mathbf{Z}$ are framed in terms of risk and odds ratios, respectively, rather than difference in expectations under the linear model. Vansteelandt[2] also showed that when R impacts $\mathbf{Z}$, the results in Section 6 still apply in terms of the risk ratio with count outcomes under the loglinear model.

However, when R impacts $\mathbf{Z}$ under the logistic model for binary outcomes, Vansteelandt[53] showed that the results in Section 6 do not apply, because of a limitation of the logistic model called the 'lack of collapsibility'.[54] This property can be defined in terms of assessing mediation or confounding by comparing the estimated marginal main effect of the randomised intervention in the outcome model without the mediator (confounder) to the corresponding estimated average controlled direct effect adjusting for the mediator (confounder). For linear and loglinear models, a substantial difference can show mediation[14] or confounding.[54] However, under the logistic model for binary outcomes, Gail et al.[54] have shown that such a difference can occur even if there is no relationship between the mediator (or confounder) and the randomised intervention. That is, a difference between the estimated marginal and controlled log odds ratios of the randomised intervention may not be indicative of mediation in contrast to the log risk ratios or mean differences nude the loglinear and linear models, respectively.

The lack of collapsibility under the logistic model precludes the application of the results in Section 6 when R affects the post-baseline confounder, $\mathbf{Z}$. Under the LSEM model with this condition, inference is based on the average controlled direct effect of the randomised intervention in $E(\theta_{CD\ m})$, which is marginal with respect to $\mathbf{Z}$. Because of the linearity of the model in terms of effects on the outcome, estimation of the marginal causal effect is possible from conditional expectations given $\mathbf{Z}$. This is also true for risk ratios under the loglinear model. In contrast, this is not possible for logistic models, because estimation with logistic regression based on conditional odds ratios given $\mathbf{Z}$ yields marginal effects that are not odds ratios.[55] Consequently,

Vansteelandt[53] presented a causal logistic approach for estimating the average controlled direct effect as an odds ratio conditional on **Z**:

$$\frac{\Pr(Y_{1,0}=1|R=r,Z=z)/\Pr(Y_{1,0}=0|R=r,\mathbf{Z}=\mathbf{z})}{\Pr(Y_{0,0}=1|R=r,Z=z)/\Pr(Y_{0,0}=0|R=r,\mathbf{Z}=\mathbf{z})} \tag{42}$$

for all r and z rather than the marginal odds ratio:

$$\frac{\Pr(Y_{1,0}=1)/\Pr(Y_{1,0}=0)}{\Pr(Y_{0,0}=1)/\Pr(Y_{0,0}=0)}. \tag{43}$$

Once **Z** is removed from (42) as a conditioning variable by integration or summing of the conditional probabilities with respect to its distribution, the resulting odds ratio conditional on only R equals the marginal odds ratio not conditional on R because of the randomisation assumption for the randomised intervention.

Another problem due to collapsibility exists when unobserved post-randomisation factors (as opposed to observed post-randomisation factors **Z**) are accounted for in the context of Section 7.1 in the absence of ignorability assumptions for M. In this case, the RPM approach in Section 7.1 for the linear model applies to the loglinear model in terms of risk ratios, but not to the logistic model. In the context of a logistic model with just a controlled log odds ratio for the mediator (e.g. non-adherence) but assuming the controlled log odds ratio for the randomised intervention is zero (e.g. exclusion restriction), Robins and Rotnitzky[56] and Vansteelandt and Goetghebeur[57] showed how collapsibility precludes working with the class of estimating equations used for the linear and loglinear link functions. In the context of estimating average controlled direct effects of the randomised intervention and mediator, the estimating equations would to be based on a logistic model which is logit-linear only in R and **X** under $\Pr(Y_{r,m}=1|R=r,\ \mathbf{X}=\mathbf{x})$. This probability model for the estimating equations would have to be obtained from the probability model used for inference, $\Pr(Y_{r,m}=1|R=r,\ M=m,\ \mathbf{X}=\mathbf{x})$, which is logit-linear in R, M, and **X**. However, the necessary integration or summation with respect to M of a model that is logit-linear in R and M will not necessarily produce a probability model logit-linear in R without M.[55] Research is underway in resolving this issue for investigating mediation when the ignorability of M and exclusion restriction are relaxed. In contrast, the PS approach in Section 7.2 is not vulnerable to this problem under the logit link for the outcome.[48,51]

An additional issue arising from mediation analyses for binary outcomes is the scale on which average controlled direct effects are interpreted. Unlike linear models, for which there is only one scale (linear or additive) non-linear models for binary outcomes present several options for the scale of interpretation: (1) the linear scale of the latent variable underlying the binary outcome and (2) non-linear scales of the binary outcome such as the logit (log odds ratios) or probit scale. Under the latent variable interpretation of mediation, the parameters of the non-linear model, which are additive on the latent scale, are affected by the constraint that the variance of the latent variable is fixed (e.g. $\pi^2/3$ for logit and 1 for probit), in contrast to the linear model where the variance is estimated. This problem is resolved by standardising the model parameters (multiply by the standard deviation of the independent variable and divide by the fixed standard deviation of the latent variable.[14] Nonetheless, the collapsibility problem still exists for these effects standardised to the linear scale.

The addition of an interaction in (20) to the mediation model under the LSEM with or without post-baseline covariates presents a problem in terms of estimation and testing of the average controlled direct effects. It turns out that even under the sequential ignorability assumptions, there may still be in identification and bias problems, if the randomised intervention impacts the mediator, as is the case under mediation. While there has been no formal work in this area, Kraemer et al.[16] noted the problem. There is current research focusing on deriving analytically the bias with confirmation by simulations.[41]

The presence of multiple mediators has been addressed by a number of authors both under sequential ignorability and absence of interactions in the LSEM context[5,22] and without the ignorability assumption for the mediators in the causal context.[25] The challenge for all of these approaches is understanding the pathways among the multiple pathways and modelling them appropriately. Dunn and Bentall[25] used G-estimation without the weights of the procedure described above in Section 7.1.3. Hence, the resulting estimators were more variable than the corresponding instrumental variable estimators also presented by Dunn and Bentall.[25] Interactions among the mediators were also estimated and tested. Once again, these interaction tests would be biased under LSEM approach under sequential ignorability because of the relationship among the mediators. In addition to multiple mediators, Cole and Maxwell[22] also addressed multiple outcomes, leading to very complex models, thus increasing the challenge for scientists to clearly understand the mechanisms of their interventions or exposures in an epidemiological context. There is still a need for more research under both the LSEM and causal approached relaxing the mediator ignorability assumption in the context of multiple mediators and longitudinal outcomes, and also accessible programmes.

9 Results for two behavioural intervention studies

The ensuing results for the two studies are taken from Ten Have et al.,[12] who focused on estimating the average controlled direct effects of the baseline intervention and mediator. First, the descriptive statistics in Table 1 suggest similarities between the two examples in terms of the ITT comparisons of outcome but not in terms of the ITT comparison of the mediator factor. That is, the ITT contrasts for outcome and mediator are significant in both studies. Hence, an analysis of the mediation of these significant ITT effects is justified. Table 1 also indicates differences between the two examples in terms of the level of use of the mediator factor by patients and also the sign of the ITT effect on the mediator factors. Most of the depressed patients in the suicide prevention study used medication regardless of whether they were in the care manager arm or not. In contrast, in the suicide therapy study, fewer of the suicidal patients used non-study therapy in either arm, although a higher proportion of the usual care group used non-study therapy than the randomised study therapy group. Given the differences between the two examples with respect to the mediator results in Table 1, we now compare the RPM, LSEM and PS results in Table 2.

9.1 Suicide prevention study

The RPM and LSEM regression estimates of the average controlled direct effects of the randomised intervention and mediator for the suicide prevention study in Table 2 are somewhat in agreement in estimating a statistically significant average controlled direct effect of the care manager intervention on the 4-month Hamilton outcome apart from increasing anti-depressant use among the depressed patients. The estimated average controlled direct effect of this intervention under both the RPM and LSEM regression approaches is an approximate reduction of 2.5 Hamilton units. However, the

Table 1. For the suicide prevention ('prevention') and cognitive therapy ('therapy') studies, means (standard deviations in parentheses) and proportions for the Hamilton or BDI depression outcomes, respectively, and proportion of patients taking anti-depressant medication or non-study therapy, respectively, by randomised intervention group or by whether they took anti-depression medication or non-study therapy

Suicide study	Group	Hamilton	Medication
Prevention	Usual care	13.55 (8.35)	0.45
	Intervention	11.50 (7.38)	0.85
	No medication	13.14 (8.09)	
	Medication	12.23 (12.23)	
		BDI	Non-study therapy
Therapy	Usual care	19.33 (12.07)	0.25
	Study therapy	14.02 (14.77)	0.08
	No non-study therapy	17.08 (14.78)	
	Non-study therapy	15.11 (12.07)	

Table 2. For the suicide prevention study, ITT, LSEM, RPM and Principal Strata (PS) estimates are presented for the direct effects of the BHS intervention and the anti-depressant mediator. PS estimates are presented for the direct effects of the BHS intervention separately in the 'fixed mediator' principal strata groups (never-medication and always-medication). Standard errors and nominal 95% confidence intervals are in parentheses

Method	Direct effect	Mediator effect
ITT[a]	−3.12 (0.82)	
	(−4.72, −1.51)	
LSEM	−2.67 (0.89)	−1.19 (0.94)
	(−4.41, −0.93)	(−3.03, 0.65)
RPM	−2.58 (1.27)	−1.43 (2.34)
	(−5.07, −0.10)	(−6.01, 3.15)
Pricipal stratum	Direct effect	
Never-med (7%)	−8.93 (6.01)	
BHS	(−17.06, 1.37)	
Always-med (36%)	−1.94 (2.18)	
BHS	(−5.23, 1.50)	

Note: [a]The ITT effect is not an average direct effect of the intervention or mediator, but instead an unadjusted average effect of the randomised intervention.

RPM confidence interval is wider than the LSEM confidence intervals, as one would expect from the MSE results in the simulations. The significant average controlled direct effect of the presence of care manager on reducing depression could be the result of the impact of this specialist on the staff and physicians of the practices. That is, one would expect that the presence of the care manager in the

intervention practices raised the sensitivity of the staff and providers in treating depression. We also see that both the RPM and LSEM regression approaches indicate a non-significant average controlled direct effect of the mediator (anti-depressant use) on outcome.

Estimating the average controlled direct effect of the care manager intervention under the RPM approach required covariates that interact with the significant randomised intervention factor on the mediator, that is varying the compliance score-based weight element, $\eta(\mathbf{x})$. One strategy for identifying such predictors is to perform logistic regression of medication use on baseline covariates stratified by randomisation arm. Comparing these predictive relationships between the two randomisation arms, the test of the overall $\mathbf{X} \times R$ interaction on M yielded a p-value of 0.006.

Finally, with the caveat that comparisons across principal strata are confounded, the PS results in Table 2 suggest heterogeneity in the direct effects across the two NT principal strata for which the mediator is held constant. In particular, the direct effect estimate in the principal stratum that would always take mediation is much smaller than that in the principal stratum group that would never take medication, as well than the RPM and LSEM estimates. However, the very wide confidence intervals for the ITT effect under the PS approach, surround zero for the ITT effects.

9.2 Suicide therapy study

In contrast to the suicide prevention study, the RPM and LSEM regression estimates of the average controlled direct effects for the suicide therapy study in Table 3 are not in agreement, indicating possible unmeasured confounding of the LSEM results and/or a violation of the no $M \times R$, $\mathbf{X} \times R$ and $\mathbf{X} \times M$ interactions assumption for Y_{rm}. Specifically, the estimate of the average controlled direct effect of CBT under the RPM is smaller than the analogous LSEM estimate. Hence, under

Table 3. For the suicide CBT study, ITT, LSEM, RPM and Principal Strata (PS) estimates are presented for the direct effects of the CBT intervention and the non-study mediator. PS estimates are presented for the direct effects of the CBT effect separately in the 'fixed mediator' principal strata groups (never-non-study and always-non-study). Standard errors and nominal 95% confidence intervals are in parentheses

Method	Direct effect	Mediator effect
ITT[a]	−6.35 (2.53)	
	(−11.37, −1.33)	
LSEM	−6.86 (2.60)	−.05 (3.46)
	(−12.01, −1.70)	(−9.92, 3.82)
RPM	−3.93 (3.09)	14.59 (15.87)
	(−9.98, 2.12)	(−16.52, 45.69)
Principal stratum	Direct effect	
Never-non-study (66%)	−7.07 (4.44)	
CBT	(−24.51, 15.67)	
Always-non-study (6%)	−8.14 (17.79)	
CBT	(−99.57, 91.38)	

Note: [a]The ITT effect is not an average direct effect of the intervention or mediator, but instead an unadjusted average effect of the randomised intervention.

the LSEM approach there is a significant average direct effect of the CBT on the 6-month depression outcome, apart from any impact on this outcome through the use of non-study therapy, whereas the RPM approach indicates that there is not sufficient evidence for such inference. There are three possible explanations, involving the sequential ignorability and no-interaction assumptions, for this discrepancy in average direct effect estimates between the RPM and LSEM approaches: (1) confounding of the non-study therapy *versus* depression outcome relationship; (2) effect modification of the non-study therapy mediator on outcome by CBT; and (3) modification of the average direct effect of baseline CBT on outcome by baseline depression or suicide ideation. Ten Have et al.[12] provided more details on the clinical implications of these three alternative explanations of the discrepancy in average controlled direct effect estimates between the RPM and LSEM approaches.

Inferentially, the RPM and LSEM approaches also disagree with respect to the sign of the average controlled effect of non-study therapy on the depression outcome, although both approaches yielded confidence intervals surrounding zero. Moreover, the RPM-based estimate and corresponding standard error are much larger in magnitude than the analogous standard regression estimates. This result conforms to the large simulation-based MSE for θ_M in Table 1 in Ten Have et al.[12] Nonetheless, Table 1 in Ten Have et al.[12] indicated such variability in the estimate of the average controlled direct effect of the mediator does not preclude more accurate inference of the G-estimation estimate of direct effect under the structural no-interaction assumption.

While not significant under either model, the estimated conditional relationship between non-study therapy and depression under the standard approach is negative (helps reduce depression) while the corresponding average controlled direct effect under the RPM approach is positive (helps increase depression). It is possible that the negative association under the standard approach may be due to confounding by environmental stresses (family and or financial) that reduced the likelihood of use of non-study therapy and also increase depression. However, when potentially controlling for this unmeasured stress confounder under the RPM, non-study therapy increases depression because of the ineffectiveness of these therapies in dealing with acute suicidal thoughts among suicide attempters.[33] Not being vulnerable to such unmeasured confounding, the RPM approach suggests that some of the ITT effect of the baseline therapy intervention occurs by reducing the reliance on non-study therapy and thus reducing depression.

In assessing the effectiveness of the multidimensional weighting for increasing efficiency of the average controlled direct effect estimators, Ten Have et al.[12] evaluated the predictors of the mediator, taking non-study therapy, stratified by randomisation arms. The corresponding test of the overall $\mathbf{X} \times R$ interaction on M yielded a p-value of 0.59, which is much less significant than the p-value of 0.006 for the larger suicide intervention study. Nonetheless, the suicide therapy study appeared to have a wider range of estimated weight elements than did the suicide prevention study, suggesting that the weights in the therapy study were still effective in improving identifiability of the causal parameters.

A reviewer questions the strategy of choosing baseline covariates based on efficiency criteria rather than the plausibility of assumptions involving the covariates. In general, choosing covariates should be based on a conceptual model. However, in this particular case, G-estimators of the average controlled effects are still unbiased under mis-specification of the relationship between the baseline covariates and outcome. Such mis-specification does impact the efficiency of these estimators as does the specification of the covariate–treatment interactions on the mediator for weights. Further, there are a number of precedents for choosing covariates based on efficiency criteria under SNMM.[26,42,58]

Finally, in contrast to the PS results in Table 2 for the study prevention study,the PS results in Table 3 reveal little heterogeneity in the average direct effects across principal strata. In particular, the average direct effect estimates in the separate 'fixed mediator' principal strata (always and never non-study therapy strata) are similar to each other and to the average controlled direct effect estimates under the LSEM and RPM approaches. Again, these results are qualified by the fact that the confidence intervals are very wide under the PS approach, surrounding zero for the ITT effects.

10 Summary

In the context of mediation analyses for baseline randomised behavioural intervention studies, we have reviewed two causal methods and one LSEM approach to estimating average controlled direct intervention effects. Traditionally, randomised studies have become the gold standard in establishing the causal effects of interventions on outcomes by allowing us to compare experimental groups using the ITT approach, which provides unbiased estimates of the average effect of randomisation. Understanding how such interventions work is needed for making these interventions more cost effective and more robust with more heterogeneous populations than the study populations on which they were tested.[59,60] Mediation analyses may satisfy these needs. However, current standard mediation methods are not protected by randomisation against potential unmeasured confounding. Consequently, causal mediation methods such as the structural mean model and PS approaches for obtaining more accurate inference under such confounding have been proposed in recent years.[12,30,46] While these causal approaches differ in terms of controlling for the possibly confounded mediator effect while estimating the average controlled direct effect of the randomised intervention, they all make tradeoffs with the no-confounding or sequential ignorability assumption for other assumptions involving treatment heterogeneity with respect to the mediator and outcome.

The tradeoffs that are made to relax the no confounding or sequential ignorability assumption under these two approaches involve model assumptions and also requirements for baseline covariate modification of randomised intervention effects on the mediator. First, there are bias *versus* variability tradeoffs shown in the simulations of Ten Have et al.[12] The RPM was shown to exhibit more variability and less bias than the LSEM approach under unmeasured confounding of the mediator effect on outcome. Gallop et al.[32] show through simulations that the PS approach also exhibits more variability but less bias than the standard mediation approach. Such variability under the PS approach was exhibited in the empirical results presented above for the two psychiatry studies. In addition, the RPM approach exchanges the untestable sequential ignorability assumption for no-interaction assumptions among baseline covariates and the randomised intervention and mediator. The PS approach makes fewer and thus more robust no-interaction assumptions. Moreover, it provides an assessment of the no-interaction assumptions made by the RPM approach. In both of the studies presented above, there was clinical conjecture about potential unmeasured confounders that would violate the sequential ignorability assumption. However, there was also clinical weight given to interactions between baseline study interventions and follow-up non-study therapies on the follow-up depression outcome. Balancing these assumptions is a clinical judgment.

In the context of estimating the marginal effect of treatment in the presence of unmeasured confounders, the instrumental variable (IV) estimator of this effect is consistent under the no-current–treatment interaction assumption. When this assumption fails, the IV estimator is a consistent estimator of the complier average causal effect under the monotonicity assumption. In the context of mediation, VanderWeele[31] presented inferential relationships between the

controlled effects and PS effects without the no-current–treatment interaction assumption. However, in contrast to the IV case, there does not appear to be a functional relationship between the two classes of parameters with or without the no-current–treatment interaction assumption or the monotonicity assumption.

Current research[61] is focusing on assessing the structural $R \times M$, $\mathbf{X} \times R$ or $\mathbf{X} \times M$ interactions under the RPM in (37). An additional element involving $\mathbf{X}$ will be added to the weight vector for each additional structural interaction parameter based on the criteria of Robins et al.[45] The difficulty of testing these structural interactions arises because $\mathbf{X}$ would be required to satisfy several strong constraints. For example for $R \times M$, $\mathbf{X}$ (e.g. baseline depression) would need to satisfy two conditions: (1) $\mathbf{x}$ leads to strong interaction with R on M (i.e. variation in compliance score across $\mathbf{x}$) and (2) $\Pr(M = 1 \mid R = 1, \mathbf{X})$ is not perfectly collinear with the compliance score. For assessing the $R \times \mathbf{X}$ interaction, condition 2) would need to be that $\mathbf{X}$ itself is not perfectly collinear with the compliance score. Our future research will focus on determining such baseline covariates satisfying these conditions for either of the two example studies. While the above weights yield consistent estimators under departures from sequential ignorability, they are not efficient under these departures. Additional future research will develop weights leading to consistent estimators that are also efficient under departures from sequential ignorability.

Additional extensions of these approaches to binary outcomes have been presented but not in the mediation context. The PS approach has been extended to causal odds ratios for different principal strata.[48,51] Robins and Rotnitzky[56] showed that additional unverifiable assumptions are needed for inference with causal odds ratios under the logistic SNMM. Accordingly, Vansteelandt and Goetghebeur[57] presented an approach that relies on an additional unverifiable assumption that a dose response of randomised intervention on outcome can be modelled correctly in the group that receives the randomised intervention. As a tradeoff to additional unverifiable assumptions, Ten Have et al.[62] presented an estimation method that approximates the true causal odds ratio under the logistic SNMM.

Finally, as we have noted, the three approaches discussed in this article differ in how the mediator is treated in defining direct effects. The LSEM and semi-parametric approaches require a hypothetical mechanism that fixes even uncontrollable mediation factors (e.g. medication use by patients at home) at a given level when specifying direct effects of the randomised intervention. In contrast, the PS approach resolves this situation by forming the principal strata on the basis of each participant's potential mediator behaviour and then only focuses on those strata for whom participants would exhibit the same mediator behaviour regardless of the randomised intervention.

Acknowledgements

The authors thank Dylan Small, Michael Elliott, Rongmei Zhang, Kevin Lynch and Mark Cary. Funding was provided by NIMH grants: R01-MH078016, R01-MH61892 and R01-CA095415.

References

1. van der Laan M and Petersen M. Direct effect models. *Int J Biostat* 2008; **4**(1): Article 23.
2. Vansteelandt S. Estimating direct effects in cohort and case-control studies. *Epidemiology* 2009; **20**(6): 851–860.
3. Bruce M, Ten Have T, Reynolds C, et al. A randomized trial to reduce suicidal ideation and depressive symptoms in depressed older primary care patients: The PROSPECT study. *J Am Med Assoc* 2004; **291**: 1081–1091.
4. Pearl J. Direct and indirect effects. In: Besnard P and Hanks S (eds) *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. San Francisco: Morgan Kaufmann, 2001, pp.411–420.

5. Robins J. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green P, Hjort N and Richardson S (eds) *In Highly structured stochastic systems*. New York: Oxford University Press, 2003, pp.70–81.
6. Baron RM and Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 1986; **51**: 1173–1182.
7. VanderWeele TJ and Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface* 2009; **2**: 457–468.
8. Imai K, Keele L and Yamamoto T. Identification, inference, and sensitivity analysis for causal mediation effects. *Stat sci* 2010; **25**(1): 51–57.
9. Rubin D. Estimating causal effects of treatment in randomised and nonrandomised studies. *J Educ Psychol* 1974; **66**: 688–701.
10. Holland P. Statistics and causal inference. *J Am Stat Assoc* 1986; **81**: 945–960.
11. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Translated by D.M. Dabrowska and edited by T.P. Speed (1990). *Stat Sci* 1923; **5**: 465–472.
12. Ten Have T, Joffe M, Lynch K, Maisto S, Brown G and Beck A. Causal mediation analyses with rank preserving models. *Biometrics* 2007; **63**: 926–934.
13. Judd CM and Kenny DA. Process analysis: Estimating mediation in treatment evaluations. *Eval Rev* 1981; **5**: 602–619.
14. MacKinnon DP, Lockwood CM, Hoffman JM, West SG and Sheets V. A comparison of methods to test mediation and other intervening variable effects. *Psychol Meth* 2002; **7**: 83–104.
15. Kraemer H, Stice E, Kazdin A, Offord D and Kupfer D. How do risk factors work together? Mediators, Moderators, and Independent, Overlapping, and Proxy Risk Factors. *Am J Psychiatry* 2001; **158**: 848–856.
16. Kraemer H, Wilson G and Fairburn C. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry* 2002; **59**: 877–883.
17. Gollob HF and Reichardt CS. Taking account of time lags in causal models. *Child Dev* 1987; **58**: 80–92.
18. Gollob HF. Interpreting and estimating indirect effects assuming time lags really matter. In: Collins LM and Horn JL (eds) *Best methods for the analysis of change: Recent advances, unanswered questions, future directions*. USA: American Psychological Association, 1991, pp.243–259.
19. Krull JL and MacKinnon DP. Multilevel mediation modelling in group-based intervention studies. *Eval Rev* 1999; **23**: 418–444.
20. Krull JL and MacKinnon DP. Multilevel modelling of individual and group level mediated effects. *Multivariate Behav Res* 2001; **36**: 249–277.
21. MacKinnon DP and Dwyer JH. Estimating mediated effects in prevention studies. *Eval Rev* 1993; **17**: 144–158.
22. Cole D and Maxwell S. Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. *J Abnorm Psychol* 2003; **112**: 558–577.
23. Kenny D, Korchmaros J and Bolger N. Lower level mediation in multi-level models. *Psychol Methods* 2003; **8**: 115–128.
24. Sobel ME. What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference. *J Am Stat Assoc* 2006; **101**: 1398–1407.
25. Dunn G and Bentall R. Modeling treatment-effect heterogeneity in randomised controlled trials of complex interventions (psychological treatments). *Stat Med* 2007; **26**: 4719–4745.
26. Robins J. Correcting for non-compliance in randomised trials using structural nested mean models. *Commun Stat Theory Methods* 1994; **23**: 2379–2412.
27. Hernan M and Robins J. Instruments for causal inference: an epidemiologists dream? *Epidemiology* 2006; **17**: 360–372.
28. Ten Have T, Elliott M, Joffe M, Zanutto E and Datto C. Causal models for randomised physician encouragement trials in treating primary care depression. *J Am Stat Assoc* 2004; **99**: 8–16.
29. Frangakis C and Rubin D. Principal stratification in causal inference. *Biometrics* 2002; **58**: 21–29.
30. Rubin D. Direct and indirect causal effects via potential outcomes. *Scand J Stat* 2004; **31**: 161–170.
31. VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Stat Probab Lett* 2008; **78**: 2957–2962.
32. Gallop R, Small D, Lin J, Elliott M, Joffe M and Ten Have T. Mediation analysis with principal stratification. *Stat Med* 2009; **28**: 1108–1130.
33. Brown G, Ten Have T, Henriques G, Xie SX, Hollander E, J and Beck AT. Cognitive therapy for the prevention of suicide attempts: a randomized controlled trial. *J Am Med Assoc* 2005; **294**: 2847–2848.
34. Robins J and Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; **3**: 142–155.
35. Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *J R Stat Soc, Ser B* 2007; **69**: 199–215.
36. Rubin D. Statistics and causal inference: comment: which ifs have causal answers. *J Am Stat Assoc* 1986; **81**: 961–962.
37. Hafeman DM and VanderWeele TJ. Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* 2010; (In press).
38. Hafeman DM. 'Proportion Explained': A causal interpretation for standard measures of indirect effect? *Am J Epidemiol* 2009; **170**: 1443–1448.
39. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* 2009; 18–26.
40. Joffe M and Greene T. Related causal frameworks for surrogate outcomes. *Biometrics* 2009; **65**: 530–538.
41. Zhang R and Ten Have T. Post-randomisation interaction analyses in clinical trial with standard regression, Unpublished manuscript, 2009.
42. Goetghebeur E and Lapp K. The effect of treatment compliance in a placebocontrolled trial: regression with unpaired data. *Appl Stat* 1997; **46**: 351–364.
43. Joffe M, Small D and Hsu C. Defining and estimating intervention effects for groups who will develop an auxiliary outcome. *Stat Sci* 2007; **22**: 7497.
44. Joffe M, Hoover D, Jacobson L, et al. Estimating the effect of Ziduvodine on Kaposi's sarcoma from observational data using a rank preserving failure time model. *Stat Med* 1998; **17**: 1073–1102.
45. Robins J, Blevins D, Ritter G and Wulfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 1992; **3**: 319–336.
46. Mealli F, Imbens G, Ferro S and Biggeri A. Analyzing a randomised trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* 2004; **5**: 207–222.
47. Angrist J, Imbens G and Rubin D. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996; **91**: 444–455.

48. Hirano K, Imbens G, Rubin D and Zhou X. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* 2000; **1**: 69–88.
49. Frangakis C, Rubin D and Zhao X-H. Clustered encouragement designs with individual noncompliance: Bayesian inference with randomisation, and application to advance directive forms. *Biostatistics* 2002; **3**: 147–164.
50. Imbens G and Rubin D. Bayesian inference for causal effects in randomised experiments with noncompliance. *Ann Stat* 1997; **25**: 305–327.
51. Frangakis C, Brookmeyer R, Varadhan R, Safaeian M, Vlahov D and Strathdee S. Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a Needle Exchange Program. *J Am Stat Assoc* 2004; **97**: 284–292.
52. Huang B, Sivaganesan S, Succop P and Goodman E. Statistical assessment of mediational effects for logistic mediational models. *Stat Med* 2004; **23**: 2713–2728.
53. Vansteelandt S. Estimation of controlled direct effects on a dichotomous outcome using logistic structural direct effect models. *Biometrika* 2010; **1** (Accepted for publication).
54. Gail M, Wieand S and Piantados S. Biased estimates of treatment effect in randomised experiments with non-linear regressions and omitted covariates. *Biometrika* 1984; **71**: 431–444.
55. Zeger S, Liang K-Y and Albert P. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**: 1049–1060.
56. Robins J and Rotnitzky A. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika* 2005; **91**: 763–783.
57. Vansteelandt S and Goetghebeur E. Causal inference with generalized structural mean models. *J R Stat Soc, Ser B* 2003; **65**: 817–835.
58. Fischer-Lapp K and Goetghebeur E. Practical properties of some structural mean analyses of the effect of compliance in randomised trials. *Control Clin Trials* 1999; **20**: 531–546.
59. Baranowski T. Theory as mediating variables: why aren't community interventions working as desired? *Ann Epidemiol* 1997; **S7**: S89–S95.
60. Kazdin A. Mediators and mechanisms of change in psychotherapy research. *Annu Rev Clin Psychol* 2007; **3**: 1–27.
61. Faerber J, Joffe M, Brown G, Beck A and Ten Have T. A causal model for postrandomisation stratification of randomised interventions. 2009; Manuscript under preparation.
62. Ten Have T, Joffe M and Cary M. Causal logistic models for non-compliance under randomised treatment with univariate binary response. *Stat Med* 2003; **22**: 1255–1284.

Note from the Editor-in-Chief

''Sophia Rabe-Hesketh has stood down from her role of editor of Statistical Methods in Medical Research due to pressure of other work. I would like to thank Sophia for all her hard work on behalf of the journal over the last ten years and for greatly contributing to the journal's continuing success.''