Statistical Methods in Medical Research 2010; **19**: 237–270

Mediation and moderation of treatment effects in randomised controlled trials of complex interventions

Richard Emsley, Graham Dunn Health Methodology Research Group, School of Community Based Medicine, University of Manchester, UK and **Ian R White** MRC Biostatistics Unit, Cambridge, UK

Complex intervention trials should be able to answer both pragmatic and explanatory questions in order to test the theories motivating the intervention and help understand the underlying nature of the clinical problem being tested. Key to this is the estimation of direct effects of treatment and indirect effects acting through intermediate variables which are measured post-randomisation. Using psychological treatment trials as an example of complex interventions, we review statistical methods which crucially evaluate both direct and indirect effects in the presence of hidden confounding between mediator and outcome. We review the historical literature on mediation and moderation of treatment effects. We introduce two methods from within the existing causal inference literature, principal stratification and structural mean models, and demonstrate how these can be applied in a mediation context before discussing approaches and assumptions necessary for attaining identifiability of key parameters of the basic causal model. Assuming that there is modification by baseline covariates of the effect of treatment (i.e. randomisation) on the mediator (i.e. covariate by treatment interactions), but no direct effect on the outcome of these treatment by covariate interactions leads to the use of instrumental variable methods. We describe how moderation can occur through post-randomisation variables, and extend the principal stratification approach to multiple group methods with explanatory models nested within the principal strata. We illustrate the new methodology with motivating examples of randomised trials from the mental health literature.

1 Introduction

Good trials evaluating complex interventions should be able to answer both pragmatic and explanatory questions. As well as asking 'Does it work?', we should also be asking 'How does it work?', 'What components are responsible for efficacy?' and 'Can it be tailored to work more effectively with particular types of patient?'.[1–4] At its best, the complex intervention trial will be a sophisticated clinical experiment designed to test the theories motivating the intervention and will also help understand the underlying nature of the clinical problem being treated.

In this review we focus on psychological treatment trials as an exemplar of complex intervention trials. Psychological treatment trials almost always involve the collection of

Address for correspondence: Graham Dunn, Health Methodology Research Group, University Place (1st Floor, Block 3), Oxford Road, Manchester, M13 9PL, UK.
E-mail: graham.dunn@manchester.ac.uk

10.1177/0962280209105014

multivariate outcomes. Rarely is it satisfactory to insist that there should be one simple primary outcome. Although these trials may be large, it would be a mistake to routinely aim to make them simple.

The term 'mediator' is commonly used for a variable on a causal pathway, and 'moderator' for a variable which modifies the strength of part or all of a causal pathway.[5] Complex interventions have, by definition, multiple components and are therefore characterised by complex treatment effect mechanisms with multiple mediators, with the possibility of moderators such as the background characteristics and environment of the patient. It is important that these mediating and moderating mechanisms are investigated as a major component of the analysis of a randomised trial. However, investigators should be aware of the pitfalls of using over-simplified methods of analysis. Simple, naïve approaches (equivalent to correlating intermediate and final outcomes) are very likely to be invalid because of 'hidden confounding' caused by selection effects. The aim of this review is to describe ways to investigate mediation and the sources of treatment effect heterogeneity in the presence of hidden confounding.

We start by describing the ideas of mediation and moderation. We then describe two motivating examples of randomised trials from the mental health literature – one aimed at suicide prevention in elderly patients suffering from depression,[6] and the other evaluating psychological interventions in recent onset psychosis.[7] We then give a brief historical survey of the theoretical work in this area and explain what hidden confounding is and why it invalidates the use of simple regression models. This is followed by a description of the relevant notation and demonstration of the decomposition of the total causal effect of treatment into direct and indirect effects in the case of a single putative mediator. We introduce principal stratification and structural mean models, and discuss approaches for attaining identifiability of the key parameters of the basic causal model. We briefly discuss moderation of treatment effects by baseline covariates and develop this idea to the evaluation of treatment effect heterogeneity due to patient characteristics, such as therapeutic alliance, that cannot be observed prior to treatment allocation (randomisation). A key component of this development is the use of explanatory models nested within latent classes (e.g. principal strata). Methods will be illustrated through analysis of data from the two randomised trials.

2 A brief introduction to mediation and moderation

We start with a trial in which there are no measured baseline covariates. Consider the simple directed or causal inference graph (path diagram) given in Figure 1(a). In such a graph, each arrow represents an assumed *causal* influence of one variable on another. Randomised treatment allocation (Z) has an effect on an intermediate outcome (M) which, in turn, has an effect on the final outcome, Y. There is also a direct effect of Z on Y. That part of the influence of Z on Y that is explained by the effect of Z on M is an indirect or mediated effect. The intermediate variable, M, is a treatment effect mediator. The key thing to remember is that Figure 1(a) is representing *structural* or causal relationships, not merely patterns of association. The effect of Z on M is the effect of manipulating Z – i.e. setting Z to equal a particular value z (set$(Z = z)$ or do$(Z = z)$, using the terminology of Pearl).[8] Similarly, the effect of M on Y is the effect

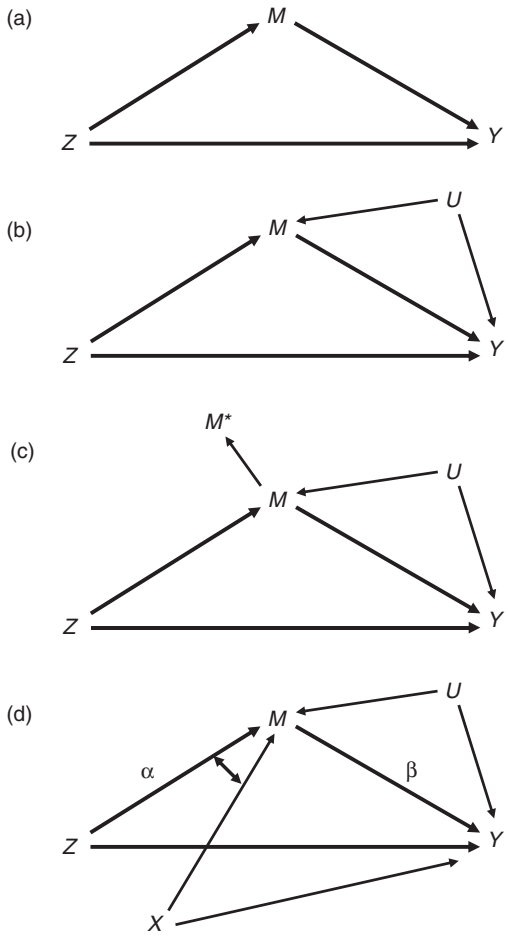

Figure 1 Possible causal path diagrams relating randomised treatment allocation (Z) to an intermediate outcome (M) and a final outcome (Y).

of manipulating M (do($M = m$)) on the outcome Y. It is not necessarily the same as the observed association between M and Y given an observed value of the mediator that has not been manipulated by the investigator.

The important skill that an investigator needs in interpreting directed graphs such as Figure 1(a) is to automatically think 'What vital component might be missing?' or 'What's not in the graph?' In our experimental set-up (the RCT), we are able to manipulate Z through random allocation (and by implication we can assume that there are no confounders of the effects of Z on either M or Y). But, typically, we have no control over either M or Y (they are *both*, in fact, outcomes of randomisation). So, there may be unobserved variables, other than treatment (Z), that influence both M and Y. Let these unobserved influences be represented by the variable U. The directed graph for this situation is shown in Figure 1(b). Let us further suppose that we cannot measure M

directly (i.e. without error), but we have an error-prone proxy, M^*. The corresponding graph is now Figure 1(c).

Let us assume that a data analyst uses simple (multiple) linear regression or structural equation modelling approaches to estimate the size of the effects illustrated by Figure 1(a). Can they interpret the resulting regression coefficients as causal effects? Yes, *if* the model represented by Figure 1(a) is the correct one. But if either Figure 1(b) or 1(c), or a more complex model, is correct then a naïve analysis based on Figure 1(a) will lead to invalid results.

Let us finally assume that we have measured an important baseline covariate, X. Suppose the effect of Z on M is influenced by the value of X. The covariate X is said to be a moderator of the effect of Z on M. In addition, X itself is assumed to influence the values of M and to *directly* influence the values of Y, but there are no covariate by treatment interactions for these components of the model. The resulting graph (assuming M is measured without error) is given in Figure 1(d). By convention, a graph with multiple arrows pointing at a single variable allows for interactions between the causal variables. We depart from this convention and indicate the interaction between X and Z on M by a double-headed arrow. Again, when interpreting Figure 1(d) we should be carefully considering what paths are missing as well as the ones that are drawn. The missing paths are indicative of some of the vital assumptions on which any valid analysis might be made.

3 Motivating examples

3.1 PROSPECT (a suicide prevention trial)

PROSPECT (Prevention of Suicide in Primary Care Elderly: Collaborative Trial) was a multi-site prospective, randomised trial designed to evaluate the impact of a primary care-based intervention on reducing major risk factors (including depression) for suicide in later life.[6] Participants were recruited from 20 primary care practices in New York City, Philidelphia and Pittsburgh regions. Ten pairs of practices were matched by region (urban vs suburban/rural), affiliation, size and population type. Within these 10 pairs, practices were randomly allocated to one of two conditions by a flip of a coin. The two conditions were either (a) an intervention based on treatment guidelines tailored for the elderly with care management or (b) treatment as usual. Bruce *et al.*[6] reported an intention-to-treat (ITT) analysis for a cohort of participants recognised as being depressed at the time of randomisation. Data from this trial have also been analysed in detail by Ten Have *et al.*,[9] Bellamy *et al.*,[10] Lynch *et al.*[11] and by Gallop *et al.*[12] in a series of papers developing and illustrating the estimation of direct and indirect treatment effects in randomised controlled trials in the presence of possible hidden confounding between the intermediate and the final outcome. An intermediate outcome (putative mediator) in the PROSPECT trial was whether the trial participant adhered to antidepressant medication during the period following allocation of the intervention. The question here is whether changes in medication adherence following the intervention might explain some or all of the observed (ITT) effects on clinical outcome. So, the focus is on the estimation of direct effects of the intervention.

Data from this trial are available on the *Biometrics* website (http://www.biometrics.tibs.org/datasets/060225CF_biomweb.zip) as supplementary material to the paper by Ten Have *et al.*[9] and comprises information on the 297 depressed elderly trial participants with complete outcome data (here the Hamilton Depression Rating Scale (HDRS)[13] score at 4 months after randomisation – the variable *hamda*). Here we use variable labels as provided in the *Biometrics* file. The baseline covariates are site (used in our analyses as the categorical factor *site*, or as two dummy variables, *s*1 and *s*2), previous use of medication (*scr*01), use of antidepressants at the time of the baseline assessment (*cad*1 – scored from 0 to 5), a dichotomised measure of suicidal ideation at baseline (*ssix*01) – based on the Scale for Suicidal Ideation (SSI),[14] and the Hamilton Depression Rating Scale total at baseline (*hamda*1). Medication adherence following treatment allocation (*interven*) is recorded by the binary variable *amedx*. Table 1 summarises the data. There appears to be a beneficial effect of the intervention on the 4-month HDRS score – could this be explained by post-randomisation changes in adherence to medication? In our analyses reported below, like those of previous authors, we make no attempt to allow for the clustering of the data within primary care practices.

3.2 SoCRATES (a schizophrenia psychotherapy trial)

The SoCRATES (Study of Cognitive Re-Alignment Therapy in Early Schizophrenia) trial was designed to evaluate the effects of cognitive behaviour therapy and supportive counselling on the outcomes of an early episode of schizophrenia. This was a multicentre prospective, rater-blind, randomised, controlled trial with planned follow-up assessments at regular intervals up to 18-months after randomisation. Participants were allocated to one of three conditions: cognitive behaviour therapy (CBT) in addition to treatment as usual (TAU), supportive counselling (SC) and TAU, or TAU alone.[7,15] Recruitment and randomisation was within three catchment areas (treatment centres): Liverpool, Manchester and Nottinghamshire. In summary, 101 participants were allocated to CBT + TAU, 106 to SC + TAU and 102 to TAU alone. 225 participants (75% of those randomised) were interviewed at 18 month follow-up, 75 in the CBT + TAU arm, 79 in the SC + TAU arm and 71 after receiving TAU alone. The remaining participants died during the follow-up period (7), withdrew consent (4) or were lost (73).

Post-randomisation variables that have a potential explanatory role in exploring the therapeutic effects include the total number of sessions of therapy actually attended and the quality or strength of the therapeutic alliance. Therapeutic alliance is a general term for a variety of therapist–client interactional and relational factors which operate in the delivery of treatment, and contains two distinct factors, a personal alliance based on the interpersonal relationship and a task related alliance based on factors of the treatment.[16] Therapeutic alliance was measured at the fourth session of therapy, early in the time-course of the intervention, but not too early to assess the development of the relationship between therapist and patient. (The alliance was also assessed at the tenth session, but we will not pursue this added complication here). The strength of the therapeutic alliance was measured in SoCRATES using two different methods, but here we report the results from an Anglicised and simplified version of the short 12-item patient-completed version of the CALPAS (California Therapeutic Alliance Scales).[17] Total CALPAS scores (ranging from 0, indicating low alliance, to 7, indicating high

Table 1 Summary statistics from the suicide prevention trial (PROSPECT)

	Site 1 (s1 = 0, s2 = 0)		Site 2 (s1 = 1, s2 = 0)		Site 3 (s1 = 0, s2 = 1)	
	Control ($N = 53$)	Intervention ($N = 53$)	Control ($N = 57$)	Intervention ($N = 54$)	Control ($N = 42$)	Intervention ($N = 38$)
Baseline characteristics: number (%)						
Antidepressant use, *cad*1 (>0)	22 (41.5)	18 (34.0)	25 (43.9)	25 (46.3)	25 (59.5)	21 (55.3)
Previous medication (*scr*01)	27 (50.9)	24 (45.3)	25 (43.9)	28 (51.9)*	29 (69.1)	20 (52.6)
Suicidal ideation (*ssix*01)	9 (17.0)	13 (24.5)	12 (21.1)	18 (33.3)	13 (31.0)	16 (42.1)
Post-randomisation adherence to antidepressant medication: number (%)						
amedx	20 (37.7)	44 (83.0)	19 (33.3)	45 (83.3)	30 (71.4)	34 (89.5)
Depression scores: mean (SD)						
At baseline (*hamda*1)	16.48 (5.33)	18.11 (6.15)	17.25 (5.26)	19.87 (6.40)	18.62 (6.32)	18.74 (5.85)
At follow-up (*hamda*)	13.42 (8.12)	11.98 (7.75)	14.10 (8.55)	12.12 (7.29)	12.98 (8.53)	9.97 (6.92)

*One missing observation.

alliance) were used in some of the analyses reported below, but we also use a binary alliance variable (1 if CALPAS score $\geqslant 5$, otherwise 0). A total of 182 (88.3%) out of 206 patients in the treated groups provided data on the number of sessions attended. Fifty-six patients from the CBT group and 58 from the SC group completed CALPAS forms at session 4 (overall 55.34%). Fewer Nottingham patients completed the scale (44.64%), compared to Manchester (65.79%) and Liverpool (52.70%). These low completion rates might be explained by patient non-compliance, therapist non-compliance or a decision by the therapist not to tax the patient.

The primary outcome measure, the PANSS (the Positive and Negative Syndromes Schedule),[18] an interview-based scale for rating 30 psychotic and non-psychotic symptoms, was administered at regular intervals by three research psychiatrists who remained blind to condition allocation. For the present purposes, only the initial (baseline) and 18-month PANSS total scores are considered. The initial PANSS score is considered as a baseline covariate in all analyses reported here. Other baseline covariates used in the analyses reported here are centre membership, the logarithm of the duration of untreated psychosis (logDUP) and years of education.

Further details and the trial outcomes have been reported elsewhere.[7,15] Briefly, from an intention-to-treat analyses of 18-month follow-up data, both psychological treatment groups had a superior outcome in terms of symptoms (as measured using the PANSS) compared to the control group. There were no differences in the effects of CBT and SC, but there was a strong centre effect, with outcomes for the psychological therapies at one of the centres (Liverpool) being significantly better than at the remaining two.

For illustrative purposes, we here ignore the distinction between CBT and SC. As indicated above, not everyone in the treated groups provided data on the number of sessions attended or on the strength of their therapeutic alliance. The analyses reported in the present article (and in previous work by Dunn and Bentall[19]) are based on all of the control group participants but only those from the CBT and SC groups who provided both a CALPAS score at the fourth session of therapy and a record of the total number of sessions attended. Note that this introduces a potential source of bias. In particular, the approach inevitably excludes the small number of participants ($n = 13$) who failed to attend sufficient sessions to have their therapeutic alliance assessed. However, our aim in using the SoCRATES data is to motivate approaches to the analysis of trials in which records of the potential explanatory covariates – particularly the amount and quality of treatment received – are complete (the more complex task of coping with several sources of missing data is beyond the scope of the present article).

Table 2 provides descriptive summaries of the SoCRATES data. It shows that there are large differences between the three centres in the mean PANSS scores at baseline, and, not surprisingly, large differences in the PANSS scores at 18-months. But there are also large differences in the treatment effects within the three centres. In centre 1 the treatment group has a mean 18-month PANSS score which is about 19 points lower (a lower score corresponds to a better outcome) than the control group; in centres 2 and 3 the differences are about $+1$ and -5 points, respectively. In the treatment groups, the number of sessions attended ranged from a minimum of 2 (presumably someone who provided a CALPAS assessment at what was intended to be the fourth session despite not attending all of the previous sessions) to a maximum of 29. The therapeutic alliance (CALPAS) scores ranged from 0 (poor) to 7 (good). Note, again, that the treated group

Table 2 Summary statistics from the SoCRATES trial

	Centre 1 (Liverpool)		Centre 2 (Manchester)		Centre 3 (Nottinghamshire)	
	Control ($N = 39$)	Treated ($N = 29$)	Control ($N = 35$)	Treated ($N = 49$)	Control ($N = 26$)	Treated ($N = 23$)
LogDUP	1.08 (0.53)	1.32 (0.53)	1.40 (0.60)	1.43 (0.63)	0.81 (0.41)	0.84 (0.38)
Years education	11.31 (1.78)	11.41 (2.65)	12.71 (2.40)	11.69 (2.02)	11.69 (2.98)	10.74 (2.80)
PANSS at baseline	80.0 (12.36)	77.7 (13.93)	97.9 (16.60)	100.5 (16.25)	84.9 (14.91)	83.4 (10.84)
PANSS at 18 months	69.5 (13.55)	50.2 (13.48)	73.2 (22.36)	74.4 (20.00)	54.5 (10.07)	49.1 (7.25)
CALPAS	–	5.73 (0.81)	–	5.07 (0.88)	–	5.15 (1.47)
Sessions	0	18.14 (3.60)	0	16.16 (4.58)	0	13.87 (4.95)
Number with high alliance N (%)*	–	23 (79.3)	–	30 (61.2)	–	13 (56.5)
Number with non-missing PANSS at 18m	23	23	25	39	21	22

*CALPAS ≥ 5. The values are given as mean (SD).

in Liverpool had the highest dose of therapy (measured by sessions attended) and also, in terms of therapeutic alliance, at least, therapy of apparently better quality (higher mean CALPAS score). Could these differences in the number and quality of sessions attended help to explain the very clear treatment group by centre interaction?

4 Brief historical survey

At present the field comprises two distinct traditions. The older and more popular approach, particularly in the social and behavioural sciences, is concerned with the estimation of direct and indirect effects through the use of path analysis (and associated regression models)[20,21] and structural equations modelling (SEM[22–26]). We will refer to this as the 'standard SEM approach'. The alternative newer approach is being developed by statisticians, econometricians and others interested in the assumptions needed for valid inferences concerning the causal effects of treatments/interventions.[9,27–34] This we will call the 'causal inference approach'.

The standard SEM approach has recently been reviewed (including a brief history) in a monograph by the psychologist MacKinnon,[35] whose methodological work has been very influential in this area (see, for example MacKinnon and Dwyer).[36] Articles by Judd and Kenny,[37] and particularly by Baron and Kenny,[5] have also been extremely influential. Baron and Kenny[5] set out three steps in the evaluation of mediation through the use of appropriate linear regression models: (1) demonstrate that treatment, Z, has an effect on the outcome, Y, (2) demonstrate that treatment, Z, has an effect on the putative mediator, M and (3) demonstrate that the mediator, M, has an effect on the outcome, Y, after controlling for treatment, Z.

Many authors, including MacKinnon,[35] have argued that the first step is not necessary. It implies that the evaluation of mediation is only of value when we have a statistically significant intention-to-treat effect on the final clinical outcome. However, analysis of mediation might also tell us why a trial result is negative. Is it because the intervention has failed to shift the mediator, or has the mediator failed to influence the outcome? Or is there a harmful direct effect of the intervention that counterbalances the benefits attained via the mediator?

Here we concentrate on steps (2) and (3). These are the regressions used to estimate and evaluate the direct and indirect (mediated) effects of Z on Y. Typically, both steps in practice will include the effects of baseline covariates, $\boldsymbol{X}$. The validity of this regression approach is dependent on the following assumptions:[35] (a) the correct functional forms (e.g. linearity) for the effect of treatment on the mediator and for the joint effects of treatment and mediator on the outcome, (b) no omitted influences (in the context of a randomised intervention, we are assuming that, conditional on the measured covariates, there is no residual or hidden confounding between mediator and outcome) and (c) the treatment, mediator and outcome are reliable and valid measures. A further assumption, that there is no co-variation between the equation errors and the explanatory variables and no co-variation between the equation errors from the two regression models, is equivalent to assumption (b). Assumption (b) is referred to as sequential ignorability in the recent statistical literature.[38–40]

Let's briefly consider *complete mediation*, where there is no direct effect of treatment on outcome. In the absence of any hidden confounding we have conditional independence between treatment and outcome: $Z \perp\!\!\!\perp Y|M, \boldsymbol{X}$ (here we use the symbol '$\perp\!\!\!\perp$' to mean 'is statistically independent of'). Now, if we have a source of hidden confounding, U, complete mediation implies $Z \perp\!\!\!\perp Y|M, \boldsymbol{X}, U$. Note that in the presence of U it is *not* true that $Z \perp\!\!\!\perp Y|M, \boldsymbol{X}$. Examination of partial correlations or the equivalent partial regression coefficients, ignoring U, will lead us astray. Similarly, hidden confounding caused by U will lead investigators astray in using regression or structural equation modelling to assess incomplete mediation. Their estimated regression coefficients will be biased. The likely presence of hidden confounding, U, is the reason why the standard SEM approach has doubtful validity.

In a randomised controlled trial, the mediator and the final clinical outcome are both outcomes of randomisation. The standard SEM approach involves controlling for the mediator (the intermediate outcome) when evaluating the direct effects of randomisation on the final outcome. The potential pitfalls of controlling for post-randomisation variables have been recognised for many years.[41] In the context of the estimation and testing of direct and indirect effects, there are several powerful critiques of the standard methods.[33,38,41–45] But it is worth noting at this point that it is not structural equation modelling (SEM) as a general technique that is necessarily at fault but that the users of the methodology are frequently fitting the wrong models.[46] Subject to solving the problems of identification, it is possible to use SEM methodology in an appropriate way (see below).

The first rigorous description of the problems arising in the estimation of direct and indirect effects appears to be that provided by Robins and Greenland.[38] A thorough exposition has also been provided by Pearl and his colleagues.[8,47,48] For other important contributions, see also Robins,[49] Tritchler,[45] Rubin,[50] Lauritzen,[51] Kaufman, *et al.*,[43] Petersen *et al.*,[52] Geneletti[53] and Goetgeluk *et al.*[54] We now review these proposals.

One way round the hidden confounding problem is to assume *a priori* that there is no direct effect of treatment (i.e. complete mediation). This leads naturally to the use of instrumental variable methods with randomisation as the instrument. Briefly, in a standard regression model, if an explanatory variable is correlated with the error term (known as endogeneity) its coefficient cannot be estimated unbiasedly. An instrumental variable (IV) is a variable that does not appear in the model, is uncorrelated with the error term and is correlated with the endogenous explanatory variable; randomisation, where available, often satisfies this criteria. A two-stage least squares (2SLS) procedure can then be applied to estimate the coefficient. At its simplest, the first stage involves using a simple linear regression of the endogenous variable on the instrument and saving the predicted values. In the second stage the outcome is then regressed on the predicted values, with the latter regression coefficient being the required estimate of the coefficient. This procedure is routinely used by econometricians and further details including the derivation of the standard errors are found in standard econometric texts such as Wooldridge.[55]

Early examples (in the context of randomised encouragement designs) are provided by Holland[33] and by Permutt and Hebel.[56] Examples of the use of instrumental variable methods to look at the effect of all-or-none compliance with randomised treatment allocation can be found in Bloom,[57] Newcombe,[58] Sommer and Zeger[59] and Follman;[60]

key theoretical discussion being provided by Imbens and Angrist[61] and Angrist *et al.*[62] (see also Baker and Lindeman).[63] The work of Angrist *et al.* led directly to the idea of principal stratification.[64] Reviews of this area of work can be found in a 2005 issue of the present journal.[65–68] In the context of quantitative measures of treatment compliance (i.e. dose–response effects), the work of Angrist and Imbens[69] is of direct relevance. The work by Robins[70] and by Goetghebeur *et al.*[71–73] on structural nested mean models and structural mean models, respectively, has been extremely influential. Maracy and Dunn[74] have recently discussed the use of instrumental variable methods to evaluate dose–response relationships in psychological treatment trials. Extensions of this work to deal with post-randomisation influences on dose–response effects can be found in Fisher and Goetghebeur[75] and in Dunn and Bentall.[19]

Evaluating both direct and indirect effects in the presence of hidden confounding between mediator and outcome is the main concern of the present review. Of related interest is the pioneering work of Gennetian and colleagues[76,77] on the use of instrumental variable methods to look at the joint effects of two or more putative mediators (we will return to this problem later in the review). Identification of the causal parameters is the major challenge.[78] Here, we are mainly dependent on the use of baseline covariates that are good predictors of an assumed heterogeneous effect of treatment allocation (i.e. randomisation) on levels of the mediator (i.e. moderators of the effect of treatment on the putative mediator). Ten Have *et al.*[9] have recently used G-estimation methods to solve the problem (see also Bellamy *et al.*[10] and Lynch *et al.*)[11] Here we observe treatment-free outcomes in those randomised to the control group and if we can deduct the effect of treatment from each of the participants allocated to the treatment group to obtain their treatment-free outcomes, then we would expect treatment-free outcome to be independent of randomisation. In essence, G-estimation is a means of finding a treatment effect estimate that makes treatment-free outcome independent of randomisation. Methods based on 2SLS and extensions of the G-estimation algorithms of Fischer-Lapp and Goetghebeur[72] have been described by Dunn and Bentall.[19] Albert[79] also used 2SLS estimation. The application of principal stratification (via maximum likelihood estimation or fully Bayesian methods) to the evaluation of direct and mediated effects in RCTs has been described by Bellamy *et al.*,[10] Lynch *et al.*,[11] Gallop *et al.*,[12] and Jo.[80] These articles also discuss the potential equivalence of the various methodological approaches.

A parallel methodological literature concerns the evaluation of surrogate outcomes (Prentice[81] and Weir and Walley).[82] Surrogate outcomes are closely related to mediating variables: a mediator may function as a good surrogate for the final outcome, and correspondingly a good surrogate may be a mediator of a treatment effect. Much of this literature ignores the problem of hidden confounding, but Joffe and Greene[83] recently discussed causal frameworks for surrogate outcomes.

Moderation (treatment effect modification) is described by Baron and Kenny[5] as a treatment by covariate interaction in a linear regression or ANCOVA/ANOVA model used to explain the effect of treatment allocation on final outcome. Here the moderator is the covariate. See Aguinis[84] and Aitken and West[85] for extensive discussions of the evaluation of interaction effects in such models. In general, moderators must precede the sources of the causal effects that they are moderating. Kraemer *et al.*[1] insist that the only covariates that are eligible to be moderators of the effect of randomisation are baseline covariates measured prior to randomisation (so that they are independent of

treatment allocation). In principle, baseline covariates could influence the size of the direct treatment effect, the effect of the treatment on the level of mediator and the effect of mediator on final outcome. We refer readers to MacKinnon[35] for further details.

Contrary to the views of Kraemer *et al.*,[1] some clinically interesting questions concern the modifying effects of variables that can only be measured once treatment has been initiated (i.e. post-randomisation). Such variables include compliance with active treatment, since the causal effect of treatment allocation is likely to be different in compliers and non-compliers (and often assumed to be zero in the non-compliers), and strength of the therapeutic alliance between patient and therapist, since the causal effect of randomisation is likely to be more beneficial in the patients who develop a good therapeutic alliance.[19] For such variables, it is useful to consider the potential value if an individual were allocated to active treatment as a baseline covariate which is observed in the active treatment arm and unobserved in the control arm. Such variables define principal strata[64] in which the effects of treatment allocation can be evaluated (again, principal stratum membership is assumed to be independent of treatment allocation). In the SEM literature, interactions are frequently evaluated by testing for heterogeneity of treatment effects across groups defined by observed covariates (the 'stacked group' analyses described by MacKinnon[35] – see Section 10), under the assumption of no unmeasured confounding. Here we are concerned with the same general idea (i.e. simultaneous estimation of treatment effects in two or more *a priori* groups) but the groups are defined by principal strata (i.e. latent classes rather than directly observed groups). The present terminology is not really satisfactory for the description of this situation – but in this review we will tentatively retain the term 'moderation'. We will return to moderation by post-randomisation covariates in Section 10.

5 Notation

We randomise participants to receive treatment or to be in the control condition. For the i-th subject (observational unit) we observe the following:

Z_i – treatment group: the outcome of randomisation ($Z_i = 1$ for treatment, 0 for controls).
$\boldsymbol{X}_i' = X1_i, X2_i, \ldots, Xp_i$ – baseline covariates.
Y_i – observed outcome.
M_i – intermediate outcome that is a putative mediator of the effects of treatment on outcome (either a quantitative measure or binary).
R_i – response: missing value indicator ($R_i = 0$ if Y_i is missing, 1 if observed).

We also define the following potential (counterfactual) outcomes:

$M_i(0)$ – mediator (intermediate outcome) if randomised to the control condition.
$M_i(1)$ – mediator (intermediate outcome) if randomised to treatment.
$Y_i(z, m)$ – outcome with treatment z and level of mediator m.
$Y_i(0) = Y_i(0, M_i(0))$ – final outcome if randomised to the control condition with intermediate outcome $M_i(0)$.

$Y_i(1) = Y_i(1, M_i(1))$ – final outcome if randomised to the treatment with intermediate outcome $M_i(1)$.

In the control arm, $Y_i = Y_i(0)$ and $M_i = M_i(0)$, so $M_i(0)$ and $Y_i(0)$ are observed and $M_i(1)$ and $Y_i(1)$ are unobserved. Similarly, in the treatment arm, $M_i(0)$ and $Y_i(0)$ are unobserved and $M_i = M_i(1)$ and $Y_i = Y_i(1)$ are observed. Apart from this, we assume for the time being that there are no missing values.

6 Effect decomposition: the total, direct and indirect effects of treatment assignment (Z)

Pearl[47] gave a formal definition of the total effect decomposition into direct and indirect effects, with application to both linear and non-linear models. Sobel[78] and Albert[79] have recently described this decomposition more specifically for linear models in areas of application similar to those covered in the present review. We follow Sobel's derivations very closely. Throughout this article we use the Stable Unit Treatment Value Assumption (SUTVA – see, for example, Rubin).[28] SUTVA has two components – (a) no interference between study units (the outcome for subject i depends only on the treatment assignment for that subject and not the treatment assignment for any other subjects), and (b) consistency, which implies that the observed outcome for unit i will equal one of the potential outcomes for that unit, no matter how the treatment was received.

The total effect of randomisation (Z) on outcome (Y) for the i-th subject is

$$Y_i(1) - Y_i(0) = Y_i(1, M_i(1)) - Y_i(0, M_i(0)).$$

Similarly, the effect of Z on the intermediate outcome or mediator (M) is

$$M_i(1) - M_i(0).$$

Taking expectations over i, we define the average treatment effect on the outcome as $\tau = E[Y_i(1) - Y_i(0)]$ and the average treatment effect on the mediator as $\alpha = E[M_i(1) - M_i(0)]$.

The total effect of randomisation can be partitioned as follows:

$$Y_i(1) - Y_i(0) = \{Y_i(1, M_i(1)) - Y_i(0, M_i(1))\} + \{Y_i(0, M_i(1)) - Y_i(0, M_i(0))\}. \quad (1)$$

The first component of this decomposition is the direct effect of randomisation given $M(1)$. The second is the effect of the change in mediator if randomised to control (i.e. $Z = 0$). Similarly,

$$Y_i(1) - Y_i(0) = \{Y_i(1, M_i(0)) - Y_i(0, M_i(0))\} + \{Y_i(1, M_i(1)) - Y_i(1, M_i(0))\}. \quad (2)$$

Here, the first component of this decomposition is the direct effect of randomisation given $M(0)$ and the second is the effect of the change in mediator if randomised to receive treatment (i.e. $Z = 1$).

We define the direct effect of treatment assignment on outcome at mediator level m as $Y_i(1, m) - Y_i(0, m)$.[111] If we are prepared to assume that this does not depend on m, then for any m and m^*,

$$Y_i(1, m) - Y_i(0, m) = Y_i(1, m^*) - Y_i(0, m^*). \tag{3}$$

Equation (3) expresses the additivity (linearity) assumption of Holland[33] and the no interaction assumption of Robins.[49] It enables us to define the mean direct effect as

$$E[Y_i(1, M_i(1)) - Y_i(0, M_i(1))] = E[Y_i(1, M_i(0)) - Y_i(0, M_i(0))] = \gamma.$$

Now, define the effect of M on Y via

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0)) = \beta(M_i(1) - M_i(0)) + \varepsilon_i \tag{4}$$

where we acknowledge lack of homogeneity of treatment effects ($\sigma_\varepsilon^2 > 0$; $E(\varepsilon_i) = 0$) but we assume that $\mathrm{Cov}(\varepsilon_i, (M_i(1) - M_i(0)) = 0$ (i.e. that there is no essential heterogeneity as defined in the econometrics literature).[86–89] It follows from (1) or (2) together with (3) and (4) that

$$\begin{aligned} \tau = E[Y_i(1) - Y_i(0)] &= E[\{Y_i(1, M_i(1)) - Y_i(0, M_i(1))\}] + \beta E[M_i(1) - M_i(0)] \\ &= \gamma + \alpha\beta. \end{aligned} \tag{5}$$

This is the decomposition from the path analysis model.

7 Binary mediator and principal stratification

If the mediator is binary ($m = 0, 1$) then the linear model of Section 6 still applies, but α has the clearer interpretation as the difference between two proportions ($\mathrm{P}(M_i(1) = 1) - \mathrm{P}(M_i(0) = 1)$). There are now four distinct possibilities (classes) for the combination of $M_i(1)$ and $M_i(0)$. These four classes define principal strata[64] and are illustrated in Table 3. It is sometimes assumed that only one of classes 3 and 4 is present (the monotonicity assumption). Individuals' class membership is not known: for example, an individual with $Z_i = 1$ and $M_i = 1$ is only known to belong to class 2 or 3.

In general, the stratum-specific average treatment effects τ_1, τ_2, τ_3 and τ_4 differ. The direct effects of treatment are measured by τ_1 and τ_2.[10–12,90] Because class membership is not changed by treatment allocation (i.e. $M_i(1), M_i(0) \perp\!\!\!\perp Z_i$), the average treatment effect for the whole trial is the weighted average of the effects within strata (i.e. $\tau = \sum_j \pi_j \tau_j$).

When we have the no interaction (additivity) assumption of Equation (3), the expected stratum-specific treatment affects are those given on the far right of Table 3.[40,80,83] Under this assumption we find as before that

$$\tau = \sum_j \pi_j \tau_j = \gamma \sum_j \pi_j + (\pi_3 - \pi_4)\beta = \gamma + \alpha\beta \tag{6}$$

since $\alpha = \pi_3 - \pi_4$. We discuss estimation of this model in Section 9.

Table 3 Principal strata with a binary mediator and the no interaction assumption

Class (stratum)	$M_i(1)$	$M_i(0)$	$M_i(1)$–$M_i(0)$	Proportion	Treatment effect	Treatment effect (no interaction)
1	0	0	0	π_1	τ_1	γ
2	1	1	0	π_2	τ_2	γ
3	1	0	1	π_3	τ_3	$\gamma + \beta$
4	0	1	−1	π_4	τ_4	$\gamma - \beta$

Within each principal stratum C ($C = 1$ to 4), and given a set of measured covariates, $\boldsymbol{X}$, we have

$$Y_i(z) = \sum_k \varphi_k X_{ik} + \tau_c z + \zeta_i \tag{7}$$

Superficially, it looks very straightforward to estimate τ_c since it is an effect of randomisation, but, of course, we typically do not know to which stratum each subject belongs. The class, C, is frequently not identified. If, however, we can find baseline covariates that are good predictors of C, then we can make progress as outlined in Section 9.

8 Structural mean models

A structural mean model is a model relating the potential outcomes $Y_i(z, m)$ to one another or to $Y_i(0, 0)$. If we assume a linear model for the potential outcome $Y_i(0, 0)$ in terms of a set of measured baseline covariates, $\boldsymbol{X}_i$ (including a vector of 1s), then as in Section 6 and in Lynch *et al.*,[11] we could write

$$Y_i(z, m) = \sum_k \lambda_k X_{ik} + \beta m + \gamma z + \varepsilon_i \tag{8}$$

for all values of z and m, with ε being independent of Z but not X and M, i.e. $E[\varepsilon_i | Z = z] = 0$.

However, it is not necessary to model counterfactuals that would not have been observed under either randomisation. For example, for an individual with $M_i(0) = M_i(1) = 1$, it is not necessary to make assumptions about $Y_i(z, 0)$. Instead, we can write

$$Y_i(1) - Y_i(0) = \gamma + \beta[M_i(1) - M_i(0)] + e_i \tag{9}$$

where $E[e_i]$=0 and $\sigma_e^2 > 0$. This term allows for treatment effect heterogeneity, with $\mathrm{Cov}(e_i, (M_i(1) - M_i(0)) = 0$.

Our aim is to estimate the causal parameters β and γ. At the same time we have a model for $M(z)$ which for quantitative $M(z)$ might be

$$M_i(z) = \sum_k \psi_k X_{ik} + \alpha z + \omega_i(z). \tag{10}$$

Again, there is a random departure term, $\omega_i(z)$ with zero expectation. The valid estimation of α and ψ is very straightforward by linear regression. Estimation of β and γ is much more problematic. In the presence of hidden confounding between M and Y, these two parameters are not identified. Often, however, we can improve our ability to identify and estimate β and γ by the introduction of covariate by randomisation interactions on the post-randomisation factor into the model for $M(z)$ (Section 9).

9 Model identification and parameter estimation: utilisation of baseline covariate by randomisation interactions

Consider a binary pre-randomisation covariate, X. Here we assume X is treatment site. We make the crucial assumption that X influences the ITT effect (τ) through its effect on the level of mediation (α) but that X does not modify either the direct effect of the intervention (γ) or the effect of the mediator on the outcome (β), as illustrated in Figure 1(d).

For site 1 we have

$$\tau_1 = \gamma + \alpha_1\beta \tag{11}$$

Similarly, for site $X = 2$

$$\tau_2 = \gamma + \alpha_2\beta \tag{12}$$

Clearly,

$$\begin{aligned} \tau_1 - \tau_2 &= (\alpha_1 - \alpha_2)\beta \\ \beta &= (\tau_1 - \tau_2)/(\alpha_1 - \alpha_2) \end{aligned} \tag{13}$$

Recall that τ_1, τ_2, α_1 and α_2 may all be estimated by regressing Y and M on Z separately in each site (possibly adjusting for other covariates). So β is now identified, as is γ (by substitution back into either (11) or (12)).

If instead the baseline covariate, X, has many levels then, in general

$$\tau_x = \gamma + \alpha_x\beta \tag{14}$$

and so β and γ are, respectively, the slope and intercept of the straight line relating the ITT effect (τ) at each level of X to the effect of treatment on the mediator (α) at that level of X. This approach has much in common with the meta-analytic regression techniques

for the evaluation of surrogate outcomes.[82,83,91,92] We note, in passing, that if a proxy for the mediator (M^*) is simply the true mediator subject to random measurement error, then using M^* rather than M itself will still yield valid causal effect estimates, as will the instrumental variable methods described below.[73,83]

Because the baseline covariate is influencing the size of the effect of treatment on the mediator, we have an X by Z interaction in the structural model for $M(z)$. There is no interaction in the model for $Y(z, m)$ and therefore the interaction is an instrumental variable (its only influence on outcome is through the mediator). So, if we have baseline covariates ($X1$ and $X2$, say) then, at an individual level we can fit an equivalent instrumental variable model through the use of 2SLS.[19,79] In Stata,[93] for example, we could use the following *ivreg* command:

$$\text{ivregress } 2sls\ y\ x1\ x2\ z(m = x1z\ x2z)$$

where $x1z$ and $x2z$ are the products of x1 or of x2 and randomisation, respectively. For a binary mediator, we might wish to use a control function approach; these are an additional function which when added to the standard regression equation removes the endogeneity because they account for the correlation between the error term and the unobserved part of the outcome.[94] A typical Stata command would be:

$$\text{treatreg } y\ x1\ x2\ z,\ \text{treat}(m = x1z\ x2z)$$

An alternative is to use a G-estimation algorithm, which conceptually is based on similar ideas involving multiple instrumental variables arising from treatment-baseline covariate interactions to estimate direct and mediation effects.[9,39,72,95] Precision is improved with the use of strong baseline covariate-treatment interactions on the mediator through weights in the estimating equations.

Jo[96] investigated the model assumptions required for identifiability when using instrumental variables to adjust for non-compliance, and highlighted two alternative assumptions which allow the exclusion restriction to be relaxed when additional covariate information is available. The first is the additivity of treatment assignment effect, which states that the ITT effect on outcome is constant regardless of varying values of the covariates. The second assumption assumes constant effects of covariates, such that the effect of a covariate on outcome does not depend on the principal strata. In the context of SoCRATES, we can therefore introduce a treatment by covariate interaction in the equation for the outcome provided it is constrained to be the same for the low and high alliance principal strata.

In the principal stratification approach, estimation proceeds by specifying a full probability model. This has been developed and illustrated by several authors (see, for example)[67,96–104] in the context of non-compliance, using models relating observed and unobserved treatment compliances (principal stratum membership) to baseline covariates, and Bayesian methods or ML/EM algorithms. In the setting of Table 3, with four principal strata, we would construct multinomial logistic regression (latent class) models to predict stratum membership using information on the impact of baseline covariates, $\boldsymbol{X}$, on potential observed mediator levels which are functions of both observed and unobserved mediator variables within each randomisation group. Accordingly, this

approach also relies on the same baseline covariate-randomisation interactions as used for the instrumental variable approach. It is also possible to fit the latent class model for stratum membership and simultaneously a further regression model for the ITT effects of treatment within each of the principal strata, usually allowing for the same baseline covariates (based on Equation (7)). If we have missing outcome data (with missing outcome indicator, R, say) we can also simultaneously fit a third model predicting missing outcomes, based on the assumptions of latent ignorability, for example.[64,67,103,104] We will return to and develop models based on these methods in Section 11.

10 Moderation of treatment and mediation effects

Instead of concentrating on moderator by treatment interactions in the more familiar regression (ANCOVA) modelling tradition, we here very briefly introduce the statistical methods based on simultaneous analyses of data from multiple groups that is a key feature of the SEM literature.[105] Let us look at sex differences in intervention effects as a simple example and let's assume that we are interested in the evaluation of the ITT effect of a psychological intervention. Using standard SEM software (Mplus[106] for example) it is very straightforward to simultaneously fit separate regression/ANCOVA models to the data from men and women. All of the model parameters can be left free to vary between the two sexes (equivalent to completely separate analyses), or some can be constrained to be equal for the two groups. Typically, we would allow the intercept term (i.e. mean outcome in the control group) to differ for men and women. However, we would frequently constrain the effects of baseline covariates and the two residual variances to be equal. Our main aim is then to test the ITT effect of treatment separately for the two sexes and to evaluate (through the testing of equality constraints) whether it might differ between the two groups. If we were evaluating a simple model for the direct and indirect effects of a mediator, M, on outcome (assuming, for the time being, that we have sequential ignorability) we can specify two simultaneous models (the effect of treatment, Z, on M; and the effect of both Z and M on outcome, Y) for the two sexes separately and then ask a series of questions concerning equality constraints on the various causal parameters within these models. The effect of the treatment on the mediator might be different for men and women, for example, but the direct effects of treatment on outcome and the effect of the mediator on the outcome might both be invariant. This type of analysis is not limited to the use of a binary moderator (one might wish to simultaneously fit mediated treatment-effect models to data from several trials, for example, or treatment centres within trials) and the models themselves can be of arbitrary complexity (involving repeated measures with missing outcomes, for example). The essence of the idea is the evaluation of between-group equality constraints in models fitted to each of the groups simultaneously. The interesting challenge, in the context of the present review, is the extension of the approach to situations in which the putative moderator variables are not fully observed (principal strata). This is the rationale for the following section. The challenge is to develop viable models and estimation procedures to evaluate the joint effect of potential mediators and moderators, stressing as always, the possibility of hidden confounding.

11 Extensions to multiple group methods: explanatory models nested within principal strata

The basic idea of principal stratification is the estimation of ITT effects within principal strata. Typically we are interested in a univariate response, but in the context of CACE estimation, Jo and Muthen[100] have investigated the advantages of simultaneously estimating CACE effects for two or more different outcomes (i.e. multivariate responses). It is possible to look at multivariate binary outcomes and, of course, one of these binary outcomes might be a missing value indicator as suggested by Frangakis and Rubin[107] in their article introducing the concept of latent ignorability.

Again, in the context of CACE estimation, Jo and Muthen[100,101] have investigated the use of latent growth curve/trajectory models for longitudinal outcome data. But the idea can be generalised to other applications of principal stratification. Other multivariate models might be equally rewarding, depending on the aims of the analysis. In the context of traditional ideas of moderation by baseline factors, it is possible to investigate patterns of mediation within each class defined by the baseline factor using either a traditional SEM approach or the newer instrumental variable or estimating equations methodology described above. It is only then a small conceptual jump to consider mediation models within the classes defined by principal strata.

An accompany article in this issue by Pickles and Croudace[108] explores the use of latent class growth models in SoCRATES with univariate and bivariate outcomes. Viewing the latent classes as described by Pickles and Croudace as principal strata, these growth mixture models identify a further class of explanatory models we can incorporate within the principal stratification framework.

Returning to CACE estimation, one may be interested in looking at mediation within the compliers, with relevant exclusion restrictions in the non-compliant groups. The idea is similar to Sobel's recent suggestion of looking at complier-average mediated effects.[78] Consider the SoCRATES trial, for example. We might be interested in two groups of participants who form principal strata defined by their potential therapeutic alliance: low alliance versus high alliance which is unchanged by treatment allocation, and observed in the treatment group, but latent in the controls. As well as looking at simple ITT effects on 18-month PANSS scores within these two principal strata, we might also be interested in the indirect effects of the number of sessions attended under the exclusion restriction, checking the sensitivity of the results to assumptions concerning hidden confounding between sessions attended and the outcome. This will be the main source of illustration in Section 12.2, but we start our illustrative analyses with a more straightforward look at the suicide prevention trial data.

12 Illustrative examples

All analyses in this section were carried out using either Stata 10[93] or Mplus 5.1[106]

12.1 PROSPECT

We start with an ITT analysis, using all the baseline variables provided in the *Biometrics* website file (that is, treatment site, antidepressant use, past mediation use, suicidal

ideation and depression severity according to the HDRS). One participant appeared to have a missing value code for previous medication use and so our analysis is based on the 296 participants with complete data. We then added post-randomisation adherence to medication as a further covariate (i.e. the standard Baron and Kenny approach to mediation). The initial ITT effect estimate was −3.15 (s.e. 0.82), indicating a small but statistically significant benefit from the intervention. In the Baron and Kenny-based analysis the direct effect of the intervention was estimated as −2.66 (s.e. 0.93).

Now, instead of following the G-estimation method of Ten Have *et al.*,[9] we used a 2SLS estimator in *Stata* (see Appendix 1(a)) to estimate the direct effect of treatment on outcome. The key feature is the use of all two-way interactions of baseline covariates with randomised intervention as instrumental variables (randomisation itself is not an instrument because we wish to estimate the direct effect of the randomised intervention on outcome). The new direct effect estimate is −2.38 (s.e. 1.35). As described by Ten Have *et al.*,[9] allowing for hidden confounding appears to have had little effect, other than increasing the standard error of the estimate. The use of maximum likelihood estimation with joint probit selection (for post-randomisation medication adherence) and outcome models (with *Stata*'s *treatreg* command – see Appendix 1(b)) produced a very similar result (−2.34 with a s.e. of 1.27). Note again the use of all of the two-way baseline by intervention interactions in the probit selection model, but not in the model for the outcomes. Since we have more baseline variables than are needed to identify the model, we note that it is possible to include one baseline by intervention interaction in the model for the outcomes in order to test the assumptions; alternatively, a general model specification test is available.[109]

Finally, we have a look at principal stratification. Here we fit a finite mixture model using the *Mplus* ML/EM algorithm, with baseline covariates predicting both latent class (principal stratum) membership and outcome (see Appendix 1(c)). We constrain the within-stratum ITT estimates to conform with our basic additive model, as above. We make the assumption of monotonicity so that there are three principal strata (classes 1, 2 and 3 in Table 3): never adherers, always adherers, and those who adhere if and only if they are in the intervention group ('compliers'). There are assumed to be no participants who only adhere if allocated to the control group ('defiers'). The direct effect of the intervention across all three principal strata is estimated to be −2.62 (bootstrap s.e. 1.38). Gallop *et al.*[12] do not assume monotonicity under a Bayesisan model and provide direct effect estimates of −2.83 (2.57) in the always-takers and −9.17 (9.41) in the never-takers. The full set of results are summarised in Table 4 and for completeness, we include Ten Have *et al.*'s estimates obtained through the G-estimation.

12.2 SoCRATES

Here, we are interested in the joint effects of the strength of the therapeutic alliance as measured by CALPAS (C) and number of sessions attended (S). Following Dunn and Bentall,[19] we postulate a structural model as follows:

$$E[Y_i(1) - Y_i(0)|X_i, S_i(1) = s, S_i(0) = 0 \,\&\, C_i = c] = \beta_{ss}s + \beta_{sc}s(c - 7) \qquad (15)$$

From the nature of the design the strength of alliance, C, can only be measured in the treatment arm. We regard this alliance measure as an indicator of an underlying latent

Table 4 Results from the analysis of the suicide prevention trial (PROSPECT)

Using all covariates, cad coded 0–4		
ITT effect:	−3.15 (0.82)	
Analytical method	Direct effect, γ (s.e.)	Indirect effect, β (s.e.)
Standard regression	−2.66 (0.93)	−1.24 (1.09)
IV (*ivreg*)	−2.38 (1.35)	−1.95 (2.71)
IV (*treatreg* – ml)	−2.34 (1.27)	−2.05 (2.49)
G-estimation (from Ten Have *et al.*[9])	−2.58 (1.27)	−1.43 (2.34)
Principal stratification (with monotonicity)	−2.62 (1.38)*	−1.37 (2.97)*

*Bootstrap standard errors.

variable that is not influenced by treatment allocation (the indicator being missing in the control group). Its influence is solely as an effect modifier (moderator) – it appears in the model as a multiplicative term and there is no effect of alliance in the absence of treatment (i.e. when $s = 0$). C is entered as (c-7) so that β_s represents the causal effect of one session of treatment in patients with maximum therapeutic alliance. Equation (15) implies an exclusion restriction – the expected treatment effect being zero when no sessions are attended. The covariates ($\mathbf{X}$) are treatment centre (represented by binary dummy variables C1 and C2), baseline PANSS score, the logarithm of the duration of untreated psychosis and years of education. Identifiability of this model is achieved because both alliance and sessions attended in the treatment group are associated with the baseline covariates (implying, together with the independence of treatment effect and covariates, that the randomisation by covariate interactions can function as instruments when using a 2SLS algorithm). Here we simply compare the 2SLS estimates with those obtained through standard (OLS) regression. In both cases we carry out a complete case analysis. Full details, together with analyses allowing for missing outcome data, are provided by Dunn and Bentall.[19] The 2SLS estimates for β_s and β_{sc} are −2.40 (s.e. 0.70) and −1.28 (s.e. 0.48), respectively. The corresponding OLS estimates are −0.95 (s.e. 0.22) and −0.39 (s.e. 0.11). Note that since β_s is the effect of the number of sessions attended in patients with maximum therapeutic alliance ($C = 7$), a negative estimate indicates a beneficial treatment effect in those with maximum therapeutic alliance. There is a statistically significant negative estimate for the sessions by alliance interaction (β_{sc}), suggesting that the benefit of treatment is lower in patients with worse therapeutic alliance. At minimum alliance (C=0), the effect of increasing sessions appears detrimental rather than helpful (for example, from the 2SLS estimates, the expected treatment effect $= -2.40s + (-7X - 1.28)s = +6.55s$). Increasing the amount of treatment appears to be helpful in those patients who form a strong working alliance with their therapist but detrimental in those who do not. Bearing this preliminary conclusion in mind, we will now change approach and look at the data using models based on principal stratification.

First, we shift to the use of the binary indicator of the strength of alliance ($A = 1$ when CALPAS>5; 0 otherwise) and ignore the number of sessions. We use *Mplus* to fit a finite mixture model based on two latent classes (high and low alliance). Latent class membership is independent of treatment allocation but is predicted by treatment centre (the dummies C1 and C2), baseline PANSS score, the logarithm of untreated psychosis

Table 5 Principal stratification in SoCRATES: treatment effect modification (moderation) by therapeutic alliance (effect estimates and their bootstrapped standard errors)

	Low alliance	High alliance
Estimated ITT effect on 18m PANSS		
Missing data ignorable (MAR)	+7.50 (8.18)	−15.46 (4.61)
Missing data ignorable (MAR)	0 (constraint)	−12.73 (4.75)
Missing data latently ignorable (LI)	+6.49 (7.26)	−16.97 (5.95)
Missing data latently ignorable (LI)	0 (constraint)*	−13.50 (5.31)
Parameter estimates from the dose–response model**		
Standard SEM		
MAR		
α	14.96 (0.96)	16.91 (0.45)
β	+0.59 (0.38)	−0.75 (0.23)
α	14.94 (0.98)	16.92 (0.46)
β	0 (constraint)	−0.61 (0.23)
LI		
α	14.94 (0.95)	16.92 (0.46)
β	+0.55 (0.42)	−0.78 (0.28)
α	14.94 (0.96)	16.93 (0.46)
β	0 (constraint)*	−0.62 (0.25)
IV SEM (*i.e. correlated errors*)		
MAR		
α	14.90 (0.97)	16.95 (0.46)
β	+0.37 (0.47)	−0.80 (0.29)
α	14.84 (0.98)	16.94 (0.46)
β	0 (constraint)	−0.71 (0.26)
LI		
α	14.85 (0.98)	16.98 (0.47)
β	+0.34 (0.50)	−0.88 (0.37)
α	14.81 (0.99)	16.94 (0.46)
β	0 (constraint)*	−0.75 (0.30)

*Compound exclusion.
**Assuming that there is no direct effect of randomisation on outcome (i.e. γ=0).
α is the effect of randomisation on sessions; β is the effect of sessions on outcome.

and years of education. We simultaneously estimate the ITT effects of treatment allocation (randomisation) within each of the two principal strata, again using the above baseline variables as covariates. The *Mplus* input file is listed in Appendix 2(a). The resulting estimate is obtained assuming that missing 18-month PANSS scores are ignorable (missing at random). The results (Table 5) indicate a clear influence of strength of alliance on the effect of treatment. The ITT estimate in those with low alliance is +7.50 (s.e. 8.18) – psychological treatment apparently being detrimental, although the effect is not statistically significant – and the ITT estimate in those with high alliance is −15.46 (s.e. 4.61). If, instead of assuming that missing data are ignorable, we introduce

a further within-stratum model to allow for latent ignorability (shown commented-out in Appendix 2(a)), then the two estimates change to +6.49 (s.e. 7.26) and −16.97 (s.e. 5.95), respectively. The choice of missing data model has little impact on the estimates. Finally, we introduce a zero ITT constraint on 18-month PANSS for the participants in the low alliance group (an exclusion restriction) and, in the case assuming missing data are LI, a compound exclusion restriction (no ITT effect on either 18-month PANSS or probability of a non-missing value). The estimated ITT effect in the high alliance principal stratum increases, presumably to compensate for the imposed reduction in the ITT effect in the low alliance stratum (Table 5).

For comparison, we make the assumption of a constant effect of covariates, as outlined in Section 9. From the analysis when assuming the missing data mechanism MAR, we obtain ITT effect estimates of −6.10 (11.67) in the low alliance group, which although not significantly different from zero is in a beneficial direction as opposed to the detrimental effect seen previously. The ITT effect in the high alliance group is −20.83 (4.81), which is similar to the previous results.

Now we turn to the more interesting problem of explaining the effect of the number of sessions attended within each of the two principal strata defined by the participants' potential therapeutic alliance. This uses the decomposed linear model for the total effect (5) introduced in Section 6 which is then fitted within the two principal strata, i.e. $\tau = E[Y_i(1) - Y_i(0)] = \gamma + \alpha\beta$ where α is the effect of random allocation on the number of sessions attended, β is the effect of sessions on outcome and γ is the direct effect of randomisation. Here the within stratum model comprises two models: a regression of the number of sessions attended on randomisation and baseline covariates (but with no interaction terms) giving α, and a regression of outcome on sessions attended and the same baseline covariates (but with a zero intercept or exclusion restriction in the sessions model) giving γ and β. Following the Baron and Kenny standard approach, the residuals from these two models are initially assumed to be uncorrelated. In a more realistic model, the residuals are explicitly allowed to have non-zero correlation (i.e. relying on the use of randomised group as an instrumental variable). We also assume that missing PANSS outcome data are either ignorable or, alternatively, latently ignorable. The *Mplus* input file is given in Appendix 2(b). The results are given in Table 5.

If the strength of the therapeutic alliance is ignored (equivalent to fitting the dose–response models ignoring principal strata) the standard and instrumental variable-based estimates for the effects of sessions on outcome are −0.36 (s.e. 0.15) and −0.46 (s.e. 0.14), respectively – see also Table III of Dunn and Bentall[19] for similar results using different software (and multiple instruments). Assuming missing 18-month PANSS data are ignorable, the corresponding estimates for the low alliance participants are +0.59 (s.e. 0.38) and +0.37 (s.e. 0.47), respectively. For the high alliance participants they are −0.75 (s.e. 0.23) and −0.80 (s.e. 0.29), respectively. Again, we see that in low alliance subjects the intervention may be detrimental, but in the high alliance stratum it is beneficial. If we assume missing data are latently ignorable the results are very similar (Table 5).

Finally, we introduce a zero constraint on the effect of sessions on 18-month PANSS for the participants in the low alliance group (equivalent to an exclusion restriction) and, in the case assuming missing data are LI, a compound exclusion restriction (no effect of sessions on 18-month PANSS and no ITT effect of randomisation on the probability

of a non-missing value). Again, the constraints produce no surprises, and the details of the results are given in Table 5.

13 Concluding thoughts

Once we acknowledge that there may be hidden confounding of the effects of the putative mediator on the final outcome (i.e. post-treatment selection effects), we have to make alternative assumptions for the identification of models that include both direct and indirect effects. Randomised treatment allocation ensures that there is no confounding of the effects of the treatment itself. Treatment allocation is also independent of the baseline covariates.

The causal inference approaches (structural mean models and principal stratification) are conceptually very similar, since both allow for unmeasured confounding but differences between the two arise from exactly how we deal with the confounded mediation effect on outcome. The structural mean model approach is similar to the standard SEM approach but explicitly allows for U by utilising X and $X*Z$ interactions. The principal stratification approach controls for the confounded mediator by stratifying the population into latent subgroups (principal strata) based on the potential mediator behaviour ($M_i(1)$ and $M_i(0)$), and fitting models within these strata.

The identifying assumptions that we have utilised in our analyses are:

(a) The effect of treatment allocation on the intermediate outcome (putative mediator) is moderated by one or more baseline covariates, denoted $X*$ and
(b) The effect of the putative mediator on the final outcome is neither moderated by treatment nor the baseline covariates $X*$ (moderation by other covariates is still possible).
(c) The direct effect of treatment allocation on the final outcome is not moderated by $X*$ (moderation by other covariates is still possible).

Assumption (a) is quite easy to test (provided we have sufficient power). Assumptions (b) and (c) need thought and are not easy to verify. If the putative mediator is a biomarker that theory tells us is (or should be) on a mechanistic pathway from treatment receipt to clinical outcome, then it might be fairly convincing to assume that its effects on the final outcome will be homogeneous. Cognitive or other psychological variables that are candidate mediators might not behave so simply. If we have several baseline covariates that appear to be moderators then we might be in the lucky position of being able to relax each of the constraints on the covariate by mediator interactions and check the sensitivity of our results using a specification test.[109] Essentially, we here have access to multiple potential instrumental variables and we are checking the sensitivity of our findings to the introduction or relaxation of exclusion restrictions to evaluate whether these covariate by treatment interactions are indeed valid instruments.

The key to the successful design of studies of potential mediation appears to be finding moderators of the effects of treatment on the proposed mediator(s). Both we and other investigators have relied on haphazard differences (arising from centre effects, for example, or the variability of effects across multiple trials). But the ideal would be to have experimental control over the moderating effects of interest. This might be built into

the design by multiple randomisations to introduce interventions specifically targeted on changing the mediator(s). Gennetian and colleagues[76,77] have recently pioneered this approach in their investigation of the parallel effects of two or more mediators. Follman[110] has also discussed design issues for the evaluation of the role of the immune response in vaccine trials. Now that many of the issues concerning the use of appropriate statistical models (and the assumptions that they imply) have been resolved, perhaps the major focus of future work should be on better design.

Acknowledgements

We thank Booil Jo, Tom Ten Have and Marshall Joffe for copies of their articles in press, and Tom Ten Have for his blessing for the use of the PROSPECT data. We thank Krista Fischer, Booil Jo and Tom Ten Have for their helpful and insightful comments on earlier drafts of this manuscript. We also thank the SoCRATES team for the use of their data. All three authors are members of the UK Mental Health Research Network (MHRN) Methodology Research Group. Research funding for a project on the explanatory (causal) analysis of randomised trials of complex interventions in mental health is provided by the UK Medical Research Council (grant number G0600555). IRW is supported by MRC grant U.1052.00.006.

References

1 Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomised clinical trials. *Archives of General Psychiatry* 2002; **59**(10): 877–83.

2 Kazdin AE, Nock MK. Delineating mechanisms of change in child and adolescent therapy: methodological issues and research recommendations. *The Journal of Child Psychology and Psychiatry* 2003; **44**(8): 1116–29.

3 Oakley A, Strange V, Bonell C, Allen E, Stephenson J. Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal* 2006; **332**(7538): 413–6.

4 Green J, Dunn G. Using intervention trials in developmental psychiatry to illuminate basic science. *British Journal of Psychiatry* 2008; **192**(5): 323–5.

5 Baron RM, Kenny DA. The moderator-mediator variable distinction in social pychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 1986; **51**: 1173–82.

6 Bruce ML, Ten Have TR, Reynolds CF *et al.* Reducing suicidal ideation and depressive symptoms in depressed older primary care patients – a randomised controlled trial. *Journal of the American Medical Association* 2004; **291**(9): 1081–91.

7 Lewis S, Tarrier N, Haddock G *et al.* Randomised controlled trial of cognitive-behavioural therapy in early schizophrenia: acute-phase outcomes. *British Journal of Psychiatry* 2002; **181**: S91–7.

8 Pearl J. *Causality* (1st edn). Cambridge University Press, New York, 2000.

9 Ten Have TR, Joffe MM, Lynch KG, Brown GK, Maisto SA, Beck AT. Causal mediation analyses with rank preserving models. *Biometrics* 2007; **63**(3): 926–34.

10 Bellamy SL, Lin JY, Ten Have TR. An introduction to causal modelling in clinical trials. *Clinical Trials* 2007; **4**(1): 58–73.

11 Lynch K, Cary M, Gallop R, Ten Have TR. Causal mediation analyses for randomised trials. *Health Services and Outcomes Research Methodology* 2008; **8**: 57–76.

12 Gallop R, Small DS, Lin JY, Elliot MR, Joffe M.M., Ten Have T.R. Mediation

analysis with principal stratification. *Statistics in Medicine* 2009; **28**(7): 1108–1130.
13 Hamilton M. A rating scale for depression. *Journal of Neurology Neurosurgery and Psychiatry* 1960; **23**(1): 56–62.
14 Beck AT, Brown GK, Steer RA. Psychometric characteristics of the scale for suicide ideation with psychiatric outpatients. *Behaviour Research and Therapy* 1997; **35**(11): 1039–46.
15 Tarrier N, Lewis S, Haddock G *et al.* Cognitive-behavioural therapy in first-episode and early schizophrenia – 18-month follow-up of a randomised controlled trial. *British Journal of Psychiatry* 2004; **184**: 231–9.
16 Green J. Annotation: the therapeutic alliance – a significant but neglected variable in child mental health treatment studies. *The Journal of Child Psychology and Psychiatry* 2006; **47**(5): 425–35.
17 Gunderson JG, Frank AF, Katz HM, Vannicelli ML, Frosch JP, Knapp PH. Effects of psychotherapy in schizophrenia: II. Comparative outcome of two forms of treatment. *Schizopreria Bulletin* 1984; **10**(4): 564–98.
18 Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* 1987; **13**(2): 261–76.
19 Dunn G, Bentall R. Modelling treatment-effect heterogeneity in randomised controlled trials of complex interventions (psychological treatments). *Statistics in Medicine* 2007; **26**(26): 4719–45.
20 Wright S. Correlation and causation. *Journal of Agricultural Research* 1921; **20**: 557–85.
21 Wright S. The method of path coefficients. *Annals of Mathematical Statistics* 1934; **5**: 161–215.
22 Simon HA. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association* 1954; **49**: 467–79.
23 Duncan OD. Path analysis – sociological examples. *American Journal of Sociology* 1966; **72**(1): 1–16.
24 Blalock HM. *Causal models in the social sciences*. Aldine-Atherton, Chicago, 2008.
25 Goldberg AS. Structural equation methods in social sciences. *Econometrica* 1972; **40**(6): 979–1001.
26 Bollen K. *Structural equations with latent variables* (2nd edn). John Wiley and Sons, Inc, New York, 1989.
27 Rubin DB. Estimating causal effects of treatment in randomised and non-randomised studies. *Journal of Educational Psychology* 1974; **66**: 688–701.
28 Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 2005; **100**(469): 322–31.
29 Heckman J. Sample selection bias as a specification error. *Econometrica* 1979; **47**(1): 153–61.
30 Heckman J. Varieties of selection bias. *American Economic Review* 1990; **80**(2): 313–8.
31 Heckman JJ. Econometric causality. *International Statistical Review* 2008; **76**(1): 1–27.
32 Holland PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**(396): 945–60.
33 Holland PW. Causal inference, path analysis and recursive structural equation models (with discussion). *Sociological Methodology* 1988; **18**: 449–484.
34 Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986; **7**(**9–12**): 1393–512.
35 MacKinnon DP. *Introduction to statistical mediation analysis* (1st edn). Taylor and Francis Group, New York, 2008.
36 MacKinnon DP, Dwyer JH. Estimating mediated effects in prevention studies. *Evaluation Review* 1993; **17**(2): 144–58.
37 Judd CM, Kenny DA. Process analysis – estimating mediation in treatment evaluations. *Evaluation Review* 1981; **5**(5): 602–19.
38 Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; **3**(2): 143–55.
39 Ten Have TR, Elliott MR, Joffe M, Zanutto E, Datto C. Causal models for randomised physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association* 2004; **99**(465): 16–25.
40 Joffe MM, Small D, Hsu CY. Defining and estimating intervention effects for groups

that will develop an auxiliary outcome. *Statistical Science* 2007; **22**(1): 74–97.
41 Herting JR. Evaluating and rejecting true mediation models: A cautionary note. *Prevention Science* 2002; **3**(4): 285–9.
42 Kaufman JS, MacLehose R, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives and Innovations* 2004; **1**(4).
43 Kaufman S, Kaufman JS, MacLehose RF, Greenland S, Poole C. Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine* 2005; **24**: 1683–702.
44 Tritchler D. Explanatory analyses of randomised studies. *Biometrics* 1996; **52**(4): 1450–6.
45 Tritchler D. Reasoning about data with directed graphs. *Statistics in Medicine* 1999; **18**(16): 2067–76.
46 McDonald RP. Haldane's lungs: A case study in path analysis. *Multivariate Behavioral Research* 1997; **32**(1): 1–38.
47 Pearl J. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. 2001; 411–20.
48 Cai ZH, Kuroki M, Pearl J, Tian J. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* 2008; **64**(3): 695–701.
49 Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In Green P, Hjort N, Richardson S, eds. *Highly structured stochastic systems*. Oxford University Press, New York, 2003: 70–81.
50 Rubin DB. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* 2004; **31**(2): 161–70.
51 Lauritzen SL. Discussion on causality. *Scandinavian Journal of Statistics* 2004; **31**(2): 189–92.
52 Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology* 2006; **17**(3): 276–84.
53 Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2007; **69**: 199–215.
54 Goetgeluk S, Vansteelandt S, Goetghebeur E. Estimation of controlled direct effects. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2008; **70**: 1049–1066.
55 Wooldridge JM. *Introductory econometrics: A modern approach* (2nd edn). Thompson Learning, USA, 2003.
56 Permutt T, Hebel JR. Simultaneous-equation estimation in a clinical-trial of the effect of smoking on birth-weight. *Biometrics* 1989; **45**(2): 619–22.
57 Bloom HS. Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 1984; **8**(2): 225–46.
58 Newcombe RG. Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. *Statistics in Medicine* 1988; **7**(11): 1179–86.
59 Sommer A, Zeger SL. On estimating efficacy from clinical-trials. *Statistics in Medicine* 1991; **10**(1): 45–52.
60 Follmann DA. On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association* 2000; **95**(452): 1101–9.
61 Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica* 1994; **62**(2): 467–75.
62 Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**(434): 444–55.
63 Baker SG, Lindeman KS. The paired availability design – a proposal for evaluating epidural analgesia during labor. *Statistics in Medicine* 1994; **13**(21): 2269–78.
64 Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**(1): 21–9.
65 White IR. Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research* 2005; **14**(4): 327–47.
66 Baker SG, Kramer BS. Simple maximum likelihood estimates of efficacy in randomised trials and before-and-after studies, with implications for meta-analysis. *Statistical Methods in Medical Research* 2005; **14**(4): 349–67.
67 Dunn G, Maracy M, Tomenson B. Estimating treatment effects from

randomised clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. *Statistical Methods in Medical Research* 2005; **14**(4): 369–95.
68 Li F, Frangakis CE. Designs in partially controlled studies: Messages from a review. *Statistical Methods in Medical Research* 2005; **14**(4): 417–31.
69 Angrist JD, Imbens GW. 2-stage least-squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 1995; **90**(430): 431–42.
70 Robins JM. Correcting for non-compliance in randomised trials using structural nested mean models. *Communications in Statistics – Theory and Methods* 1994; **23**(8): 2379–412.
71 Goetghebeur E, Lapp K. The effect of treatment compliance in a placebo-controlled trial: Regression with unpaired data. *Applied Statistics-Journal of the Royal Statistical Society Series C* 1997; **46**(3): 351–64.
72 Fischer-Lapp K, Goetghebeur E. Practical properties of some structural mean analyses of the effect of compliance in randomised trials. *Controlled Clinical Trials* 1999; **20**(6): 531–46.
73 Goetghebeur E, Vansteelandt S. Structural mean models for compliance analysis in randomised clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research* 2005; **14**(4): 397–415.
74 Maracy M, Dunn G. Estimating dose-response effects in psychological treatment trials: The role of instrumental variables. *Statistical Methods in Medical Research* 2008; (In press).
75 Fischer K, Goetghebeur E. Structural mean effects of noncompliance: estimating interaction with baseline prognosis and selection effects. *Journal of the American Statistical Association* 2004; **99**(468): 918–28.
76 Gennetian LA, Morris PA, Bos JM, Bloom HS. Constructing instrumental variables from experimental data to explore how treatments produce effects. In Bloom HS, ed. *Learning more from social experiments: evolving analytic approaches* (1st edn). Russell Sage Foundation, New York, 2005: 75–114.
77 Gennetian LA, Magnuson K, Morris PA. From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology* 2008; **44**(2): 381–94.
78 Sobel ME. Identification of causal parameters in randomised studies with mediating variables. *Journal of Educational and Behavioral Statistics* 2008; **33**(2): 230–51.
79 Albert JM. Mediation analysis via potential outcomes models. *Statistics in Medicine* 2008; **27**(8): 1282–304.
80 Jo B. Causal inference in randomised experiments with mediational processes. *Psychological Methods* 2008; **13**(4): 314–36.
81 Prentice RL. Surrogate endpoints in clinical-trials – definition and operational criteria. *Statistics in Medicine* 1989; **8**(4): 431–40.
82 Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: A literature review. *Statistics in Medicine* 2006; **25**(2): 183–203.
83 Joffe M.M., Greene T. Related causal frameworks for surrogate outcomes. *Biometrics* 2008; (In press).
84 Aguinis H. *Regression analysis for categorical moderators*. Guilford, New York, 2004.
85 Aitken LS, West SG. *Multiple regression: Testing and interpreting interactions*. SAGE, Newbury Park, 1991.
86 Heckman J, Vytlacil E. Instrumental variables methods for the correlated random coefficient model – estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources* 1998; **33**(4): 974–87.
87 Heckman JJ, Urzua S, Vytlacil E. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 2006; **88**(3): 389–432.
88 Wooldridge JM. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters* 1997; **56**(2): 129–33.
89 Wooldridge JM. Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters* 2003; **79**(2): 185–91.

90 Mealli F, Rubin DB. Commentary – Assumptions allowing the estimation of direct causal effects. *Journal of Econometrics* 2003; **112**(1): 79–87.

91 Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**(17): 1965–82.

92 Burzykowski T, Molenberghs G, Buyse M. *The evaluation of surrogate endpoints*. Springer, New York, 2006.

93 *Intercooled Stata Statistical Software: Release 10.0* [computer program]. Stata Corporation, College Station, TX; 2007.

94 Florens JP, Heckman JJ, Meghir C, Vytlacil E. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica* 2008; **76**(5): 1191–206.

95 Robins JM, Greenland S. Adjusting for differential rates of prophylaxis therapy for PCP in high-dose versus low-dose AZT treatment arms in an AIDS randomised trial. *Journal of the American Statistical Association* 1994; **89**(427): 737–49.

96 Jo B. Estimation of intervention effects with noncompliance: alternative model specifications. *Journal of Educational and Behavioral Statistics* 2002; **27**(4): 385–409.

97 Imbens GW, Rubin DB. Bayesian inference for causal effects in randomised experiments with noncompliance. *Annals of Statistics* 1997; **25**(1): 305–27.

98 Imbens GW, Rubin DB. Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies* 1997; **64**(4): 555–74.

99 Little RJ, Yau LHY. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychological Methods* 1998; **3**(2): 147–59.

100 Jo B, Muthén BO. Modeling of intervention effects with noncompliance: a latent variable approach for randomised trials. In Marcoulides GA, Schumacker RE, eds. *New developments and techniques in structural equation modeling*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2001: 57–87.

101 Jo B, Muthén BO. Longitudinal studies with intervention and noncompliance: estimation of causal effects in growth mixture modeling. In Duan N, Reise S, eds. *Multilevel modeling: methodological advances, issues, and applications*. Lawrence Erlbaum Associates, Mahwah, NJ, 2002: 112–39.

102 Yau LHY, Little RJ. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* 2001; **96**(456): 1232–44.

103 Peng YH, Little RJA, Raghunathan TE. An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics* 2004; **60**(3): 598–607.

104 Mealli F, Imbens GW, Ferro S, Biggeri A. Analyzing a randomised trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* 2004; **5**(2): 207–22.

105 Croudace T, Dunn G, Pickles A. *General latent variable modelling using mplus*. Chapman and Hall, London, 2010.

106 Muthén LK, Muthén BO. *Mplus user's guide*. Muthén and Muthén, Los Angeles, CA, 1998.

107 Frangakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 1999; **86**(2): 365–79.

108 Pickles A, Croudace T. Latent mixture models for multivariate and longitudinal outcomes. *Statistical Methods in Medical Research* 2010; **19**(3): 271–89.

109 Hausman JA. Specification tests in econometrics. *Econometrica* 1978; **46**(6): 1251–71.

110 Follmann D. Augmented designs to assess immune response in vaccine trials. *Biometrics* 2006; **62**(4): 1161–9.

111 Robins JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In Glymour C, Cooper G, eds. *Computation, Causation, and Discovery*. AAAI Press/The MIT Press, Menlo Park, CA, Cambridge, MA, 1999: 349–405.

Appendix 1

(a) Using *Stata's ivreg* command: 2SLS estimation of direct and indirect effects of intervention in PROSPECT.

```
xi: ivreg hdrs4 hdrs0 cad1 ssix01 scr01 i.site i.interven (amedx = i.interven*hdrs0
i.interven*cad1 i.interven*ssix01 i.interven*scr01 i.interven*i.site)
```

(b) Using *Stata's treatreg* command: maximum likelihood estimation (with a probit first-stage selection model) of direct and indirect effects of intervention in PROSPECT.

```
xi: treatreg hdrs4 hdrs0 cad1 ssix01 scr01 i.site i.interven, treat(amedx =
i.interven*hdrs0 i.interven*cad1 i.interven*ssix01 i.interven*scr01 i.interven*i.site)
```

(c) Mplus input file for analysis of Suicide Prevention trial using three principal strata (i.e. assuming monotonicity)

```
Note: anything on a line to the right of ! is an explanatory
comment (ignored by Mplus). These comments follow the command
lines they are attempting to explain.

DATA: FILE IS suicide.raw;
VARIABLE: NAMES ARE cad1 hdrs1 ssix01 scr01 hdrs0 site
          interven amedx c1 c2 c3 s1 s2;

! s1 and s2 are binary dummy variables to indicate
! treatment site.
! c1, c2 and c3 are binary indicators for likely membership of
! the three principal strata (never on medication, always

! on medication, and on medication as a result of the
! intervention, respectively). The indicator is coded 1 if the
! relevant class membership is possible, 0 otherwise.
! Examples:
! if interven=1 and amedx=1 then c1=0, c2=1 and c3=1
! if interven=1 and amedx=0 then c1=1, c2=0 and c3=0
! if interven=0 and amedx=0 then c1=1, c2=0 and c3=1
! if interven=0 and amedx=1 then c1=0, c2=1 and c3=0

         CLASSES C(3);

! The three principal strata.

         TRAINING=C1-C3;

! Data using in determination of latent classes (see above).

         USEVARIABLES cad1 hdrs0 ssix01 scr01 hdrs1
         interven C1 C2 C3 s1 s2 cmed;
```

```
! See below for the explanation of the role of the previously
! unmentioned variable cmed.

         MISSING scr01(8);

DEFINE:     cmed=interven;

! The new variable cmed (identical to the randomised intervention) is
! used as a means of estimating the effect of medication on outcome
! (see below).

ANALYSIS: TYPE=MIXTURE;
          STARTS = 5000 10;
          ESTIMATOR=ML;
          BOOTSTRAP=250;

! Here we are fitting a finite mixture (latent class model) using an
! ml/em algorithm (5000 starts with randomly perturbed starting
! values), incorporating bootstrapping to estimate standard errors
! (250 bootstrap samples).

MODEL:
          %OVERALL%
          hdrs1 ON cad1 hhdrs0 ssix01 scr01 s1 s2 interven cmed;
          C#1 ON cad1 hdrs0 ssix01 scr01 s1 s2;
          C#2 ON cad1 hdrs0 ssix01 scr01 s1 s2;

! These are the three models applicable to all participants in the
! trial. The first line corresponds to a multiple regression model
! for the outcome to estimate the effects of the intervention,
! allowing for all baseline covariates. The parameter corresponding
! to interven is the direct effect of the intervention (common to all
! three principal strata). All parameters except for the effect of
! cmed will be constrained to be equal across the three principal
! strata. The effect of cmed is the effect of taking medication in
! those participants (the compliers) who have begun to take
! medication as a result of the intervention. The second two lines
! are for logistic regression models for principal stratum
! membership.

          %C#1%         ! Never take medication.
          [hdrs1];

! Allows the intercept for the outcome to be freely estimated in this
! class (i.e. not constrained to be equal to that for the other two
! classes).
```

```
          hdrs1 ON cmed@0;
! Short-hand version of the full multiple regression model above,
! but the parameter corresponding to the effect of medication
! being constrained to be zero (@0).

          %C#2%          ! Always take medication
          [hdrs1];
          hdrs1 ON cmed@0;

          %C#3%          ! Take medication as a result of the
                         ! intervention
           [hdrs1];
           hdrs1 ON cmed*0;

! The effect of medication in this class (i.e. the indirect effect
! of the intervention) is free to be estimated (*0, with 0
! indicating an arbitrary starting value).
```

Appendix 2

Mplus input file for analysis of SoCRATES using two principal strata

(high vs low alliance)

(a) ITT effect within strata. Missing outcome assumed to be latently ignorable (LI)

```
TITLE:    Principal stratification - SoCRATES
DATA:     FILE IS Socrates_alliance.raw;
VARIABLE: NAMES logdup pantot pant18 sessions yearsed c1 c2
          rgroup alliance resp;
          CLASSES C(2);
          CATEGORICAL alliance resp;
          USEVARIABLES logdup pantot pant18 yearsed c1 c2
          rgroup alliance resp;
          MISSING pant18(999) alliance(999);
ANALYSIS: TYPE=MIXTURE;
          STARTS = 100 10;
MODEL:    %OVERALL%
          resp ON logdup pantot yearsed c1 c2 rgroup;
          pant18 ON logdup pantot yearsed c1 c2 rgroup;
          C#1 ON logdup pantot yearsed c1 c2;

! There are three models here. The first is a logistic regression
! to predict the indicator of a non-missing outcome (resp).
! The second is a multiple regression for the outcome itself.
```

```
! The third is a logistic regression for latent class membership
! (high versus low alliance). All parameters for the missing data
! and outcome models are constrained to be equal for the two classes
! unless otherwise indicated below.

          %C#1%        ! Low Alliance
          [alliance$1@15];

! A declared threshold to force participants with recorded alliance=0
! into this class.

          [resp$1];
          resp ON rgroup*0;
          [pant18];
          pant18 ON rgroup*0;

! These statements release the equality constraints on the relevant
! model intercept terms for the effects of the randomised
! intervention.

          %C#2%        ! High alliance
          [alliance$1@-15];

! A declared threshold to force participants with recorded alliance=1
! into this class.

          [resp$1];
          resp ON rgroup*0;
          [pant18];
          pant18 ON rgroup*0;
```

(b) Dose–response models within principal strata (missing data assumed LI)

```
TITLE:    Principal stratification - SoCRATES
DATA:     FILE IS Socrates_alliance.raw;
VARIABLE: NAMES logdup pantot pant18 sessions yearsed c1 c2
          rgroup alliance resp;
          CLASSES C(2);
          CATEGORICAL alliance resp;
          USEVARIABLES logdup pantot pant18 sessions yearsed c1 c2
          rgroup alliance resp;
          MISSING pant18(999) alliance(999);

ANALYSIS: TYPE=MIXTURE MISSING;
          starts = 100 10;
```

```
          estimator=ml;
          bootstrap=250;

MODEL:

          %OVERALL%
          resp ON logdup pantot yearsed c1 c2 rgroup;
          sessions ON logdup pantot yearsed c1 c2 rgroup;
          pant18 ON sessions logdup pantot yearsed c1 c2;
          pant18 WITH sessions;
          C#1 ON logdup pantot yearsed c1 c2;

! Here we have five statements: one for the missing data
! model; another model for the effect of the intervention and
! baseline covariates on sessions attended; a model for the
! effects of sessions and baseline covariates on outcome
! (no effect of the intervention here) together with a line to
! indicate that there is correlation between the residuals from
! the sessions and outcomes models (pant18 with sessions).
! The last line is the latent class model for alliance status.

          %C#1%                          ! Low Alliance
          [alliance$1@15];
          [resp$1];
          resp ON RGROUP*0;
          [sessions];
          sessions ON rgroup*0;
          [pant18];
          pant18 ON sessions*0;

          %C#2%                          ! High alliance
          [alliance$1@-15];
          [resp$1];
          resp ON RGROUP*0;
          [sessions];
          sessions ON rgroup*0;
          [pant18];
          pant18 ON sessions*0;
```