Published in final edited form as:
Stat Med. 2009 March 30; 28(7): 1108–1130. doi:10.1002/sim.3533.

Mediation Analysis with Principal Stratification

Robert Gallop,
Department of Mathematics, Applied Statistics Program, West Chester University

Dylan Small,
Wharton School of Business, University of Pennsylvania

Julia Y. Lin,
Center for Multicultural Mental Health Research, Cambridge Health Alliance

Michael R. Elliot,
Department of Biostatistics, University of Michigan

Marshall Joffe, and
Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine

Thomas R. Ten Have
Department of Mathematics, Applied Statistics Program, West Chester University

Abstract

In assessing the mechanism of treatment efficacy in randomized clinical trials, investigators often perform mediation analyses by analyzing if the significant intent-to-treat treatment effect on outcome occurs through or around a third intermediate or mediating variable: indirect and direct effects, respectively. Standard mediation analyses assume sequential ignorability, i.e., conditional on covariates the intermediate or mediating factor is randomly assigned, as is the treatment in a randomized clinical trial. This research focuses on the application of the principal stratification approach for estimating the direct effect of a randomized treatment but without the standard sequential ignorability assumption. This approach is used to estimate the direct effect of treatment as a difference between expectations of potential outcomes within latent sub-groups of participants for whom the intermediate variable behavior would be constant, regardless of the randomized treatment assignment. Using a Bayesian estimation procedure, we also assess the sensitivity of results based on the principal stratification approach to heterogeneity of the variances among these principal strata. We assess this approach with simulations and apply it to two psychiatric examples. Both examples and the simulations indicated robustness of our findings to the homogeneous variance assumption. However, simulations showed that the magnitude of treatment effects derived under the principal stratification approach were sensitive to model mis-specification.

Keywords

Principal stratification; mediating variables; direct effects; principal strata probabilities; heterogeneous variances

Contact: Robert Gallop 610 436 2419 or E-mail: rgallop@wcupa.edu Department of Mathematics, Applied Statistics Program, West Chester University, 323B Anderson Hall, West Chester, PA 19383.

1. Introduction

We present a principal stratification model [1-3] approach for investigating whether a randomized baseline intervention effect on a continuous outcome occurs around a post-randomization binary intermediate factor (direct effect) in the context of randomized behavioral health trials. Estimating such an effect is part of a mediation analysis of whether the effect of treatment is through or around the intermediate factor. (e.g., [4-6]).

An assumption that is commonly made for standard mediation analysis methods (e.g., [4]) is the untestable, no unmeasured confounding assumption for the intermediate factor, illustrated in Figure 1. Therefore, it is assumed that there is no extraneous variable which influences both the mediator and the outcome. If such variables were present then an assessment of direct effect of treatment would require some adjustment due to the influence of the confounding variable.

This assumption is equivalent to randomization of the baseline intervention and of subsequent intermediate variables (i.e., a strong form of “sequential ignorability”; e.g., [7]). Because intermediate factors are typically not randomized, we use the Principal Stratification approach as one way of relaxing the sequential ignorability assumption by imposing model constraints.

The principal stratification (PS) model stratifies the population into latent classes or principal strata based on the potential values of the mediator variable under the randomized treatment assignment. Because these principal strata are based on potential outcomes for the mediator under different randomized treatment conditions for each individual, treatment effects on outcome within each principal strata can be interpreted causally [1,8]. In noncompliance to treatment assignment settings, PS models have been used to estimate the effects of the randomized treatment assignment on outcome (i.e., intent-to-treat (ITT) effects) within separate principal strata based on potential compliance behavior under each randomization assignment (e.g [9-10]).

In the context of mediation, Mealli and Rubin [11] and Rubin [2] noted that direct effects corresponded to the ITT effects on outcome in those principal strata for which the potential mediator level is constant when the treatment is varied. We will assess such a model with examples applied to the example datasets and with simulations. In addition, we will consider this model with heterogeneous variances to assess the robustness of the PS mediation approach to the constant variance assumption made previously (e.g., [12]).

Other causal approaches have been proposed to estimating direct effects without the sequential ignorability assumption. Ten Have et al. [6] proposed a rank preserving model with weighted G-estimation to investigate direct effects for the complete sample in contrast to within specific principal strata under consideration in this paper. Dunn et al. [13] proposed an analogous instrumental variable approach to estimating direct effects. We apply the PS approach to the datasets analyzed in Ten Have et al. [6], and make some comparisons to the results from the rank preserving model.

We consider two examples for which we want to investigate the direct effect of treatment. The first sample, a suicide therapy study, reported in Brown et al. [14] consists of 101 adults who attempted suicide. These patients were evaluated within 48 hours of the suicide attempt as well as at the conclusion of 6 months of treatment. Patients were randomized to Cognitive Behavioral Therapy (CBT) or Usual Care (TAU). In addition to receiving the counseling services available through the treatment regimen, some patients sought additional therapy outside the study. The primary outcome of interest was the depression as measured by the Beck Depression Index (BDI) at the end of the 6 months of treatment [15]. We investigate whether or not outside non-study therapy may be mediating the significant intent-to-treat effect of the CBT on 6 months depression. The second sample, a suicide prevention study, reported in Bruce et al. [16] consists of 297 patients who were randomized to family practices either consisting of depression specialist versus those family practices that do not. In addition to receiving the services provided in their family practice, some patients took medication for their depression. The primary outcome of interest for this sample is their depression as measured by the Hamilton Depression Scale (HRSD) at 4 months post randomization [17-18]. In this example, it is of interest to investigate whether or not receiving additional medication may be mediating the significant intent-to-treatment effect of the randomized depression specialist in the primary care practice.

The remainder of the paper is organized as follows. In section 2, we present the principal stratification model with particular focus on estimating direct effects within certain principal strata and also the accommodation of heterogeneous variances and inclusion of covariates. In section 3, we consider a simulation to assess the validity and robustness of our estimation strategy. In section 4, we discuss the estimation of PS model-based direct effects in certain principal strata for the two example datasets with a comparison to a standard regression approach to estimating direct effects for the whole sample, assuming sequential ignorability (e.g., [4]). We also assess sensitivity of the PS approach to homogeneous variance assumptions. In section 5 we provide some concluding remarks.

2. Principal Stratification Model

In this section, the principal stratification model estimates direct effects as intent-to-treat effects conditional on the principal strata. The model is presented first in terms of the potential outcomes approach [19-20]. Estimation is achieved in a subsequent section by translating to an equivalent model for the observed outcomes conditional on principal strata.

2.1 Notation

Notation for the observed and potential outcome below suppresses the patient index i for simplicity. The observed variables include the mediator variable, D, the randomized treatment assignment variable, R, and a vector of observed baseline covariates **X**. The mediator, D, is defined as taking the mediator (D=1) or not (D=0). The treatment assignment, R, is defined as treatment (R=1) or control (R=0).

For causal interpretations of the ITT effects within each principal strata, we define potential variables corresponding to the observed variables, D and Y, the outcome variable, under each randomization assignment. The notation D_r corresponds to the mediator value when R=r. With binary r, there are two separate potential mediator variables: D_0 and D_1. D_1 is the mediator variable that would be observed if a subject was randomized to the treatment; and D_0 is the mediator variable if the same participant were randomized to the control group. Similarly, we define the potential outcome Y_r as the outcome that would be observed if a given patient were randomized to level r. There are two separate potential outcome variables: Y_0 and Y_1. Y_1 is the outcome variable that would be observed if a subject was randomized to the treatment; and Y_0 is the outcome variable if the same participant were randomized to the control group. With these two potential outcome variables, we can define the causal contrasts for the direct effect of the baseline intervention within certain principal strata.

2.2 Principal Strata

Under the PS approach applied to the mediation context, we focus on the ITT effects in two of the four principal strata formed by the potential behavior of a binary mediator with a binary randomized treatment assignment. For the first principal stratum, compliant mediators, all subjects in this principal stratum exhibit the positive mediator behavior when assigned to treatment but do not exhibit the positive mediator behavior when assigned to the control group

(D_1 =1 and D_0 =0). The second principal stratum, always mediators, consists of subjects who regardless of treatment assignment, always exhibit the positive mediator behavior (D_1 =1 and D_0 =1). The third principal stratum, never mediators, consists of subjects who regardless of treatment assignment never exhibit the positive mediator behavior (D_1 =0 and D_0 =0). The fourth stratum, defiant mediators, is the converse of the compliant mediating stratum; therefore, it consists of subjects who when assigned to treatment do not exhibit positive mediator behavior but when assigned to the control do exhibit positive mediator behavior (D_1 =0 and D_0 =1).

For the PS model, no randomization assumptions for the mediator are required, since the potential mediator behavior under each treatment assignment regardless of actual treatment assignment identifies the principal stratum under certain model assumptions and covariate relationships (e.g., [1]). Mediation analyses are then based on the intent-to-treat analysis within the never and always mediator strata. Because the potential mediator level is constant within each of these principal strata, the separate ITT effect of treatment within each of these strata is a direct effect.

2.3 Principal Stratification Model

Inference on the direct effect of the intervention holding the mediator constant is based on the ITT effects of treatment within the always and never mediator strata. These ITT effects are defined in the following models for the outcome stratifying on the principal strata, as illustrated in Figure 2.

Figure 2 shows that for the always and never mediator strata, the pathway between treatment and outcome does not include the mediator, because the potential level of the mediator is held constant in these two strata. Given the principal strata are defined on the basis of potential mediator behavior under each randomization assignment, the ITT effects within these principal strata are causal in that they are protected against unmeasured confounding of the mediator-outcome relationship.

A separate model is specified for each pair of potential outcomes Y for each principal strata, where r takes on the values 1 for treatment and 0 for the non-treatment. Per strata, we have:

$$Y_r=\theta_{\mathrm{ITT}_t} r+\beta_t^T X+\varepsilon_{tr}. \tag{1}$$

The subscript t takes on the values 1, 2, 3, and 4 corresponding to the compliant mediating, always mediating, never mediating, and defiant mediating principal strata, respectively. As addressed by others (e.g., [1]), we note that the correlation between ε_{t0} and ε_{t1} is not identifiable from the data under the proposed model and assumptions.

As an extension, we consider modeling separate variances within each principal strata. Therefore, under this heterogeneity extension, we have:

$$\varepsilon_{tr}\sim\mathcal{N}\left(0,\sigma_t^2\right) \quad \text{for} \quad t\in\{1,2,3,4\}. \tag{2}$$

The ITT effect for the tth principal stratum for subjects with covariates ***X*** then is:

$$\theta_{\mathrm{ITT}_t}=E\left[Y_1|X,C=t\right]-E\left[Y_0|X,C=t\right], \tag{3}$$

where **X** is a vector of baseline covariates, and C=t corresponds to the tth principal stratum. The standard ITT effect for the population equals the weighted sum of the stratum-specific ITT effects across all four strata with weights corresponding to probabilities of membership in each principal stratum, π_t = P (Cr=t|**X**), where **X** are baseline covariates, such that $\sum_{t=1}^{4}\pi_t=1$. For notational simplicity, we will suppress the dependence of $\theta_{\mathrm{ITT}t|X}$ and $\pi_{t|X}$ on ***X*** hereafter. We model the π_t 's as functions of the baseline covariates with the following multinomial model:

$$\pi_t=P\left(C=t|\mathbf{X}=\mathbf{x}\right)=\frac{\exp\left(\delta_{\mathrm{t}}^{\mathrm{T}}\mathbf{x}\right)}{\sum\limits_{t=1}^{4}\exp\left(\delta_{\mathrm{t}}^{\mathrm{T}}\mathbf{x}\right)} \tag{4}$$

Note the covariates in the outcome model and the principal stratum probability do not need to coincide, but for simplicity we use the same notation. Under no specification of covariates the principal strata model is parameterized as follows:

$$\pi_t=P\left(C=t\right)=\frac{\exp\left(\delta_t\right)}{\sum\limits_{t=1}^{4}\exp\left(\delta_t\right)} \tag{5}$$

Therefore, we have:

$$\theta_{ITT}=\sum_{t=1}^{4}E\left[Y_1-Y_0|\mathbf{X}=\mathbf{x},C=t\right]\pi_t \tag{6}$$

The direct effect of treatment corresponds to weighted sum of the ITT effect across the always and never mediating strata and is computed as follows:

$$\theta_{de}=\frac{\sum\limits_{t=2}^{3}E\left[Y_1-Y_0|\mathrm{X}=\mathrm{x},C=t\right]\pi_t}{\pi_2+\pi_3}. \tag{7}$$

2.4 Assumptions

To be able to estimate the above parameters under the specified model, a number of assumptions are needed: 1) the Stable Unit Treatment Value Assumption (SUTVA); 2) randomization of treatment; and (3) baseline covariate-randomization no-interaction assumptions.

The SUTVA assumption consists of two parts [21]. First, there is a single value for each of the potential random outcome variables $\mathrm{Y_r}$ for a given patient regardless of the randomization assignment of any other patient. Notationally, this assumption implies $\mathrm{Y_r}$ is defined with scalar indices for a given patient, rather than vector indices representing baseline intervention levels for all subjects. Second, there is a single value for each potential outcome random variable $\mathrm{Y_r}$ for a given patient regardless of the method of administration of the randomized baseline intervention, such that for a given patient with observed level r for R,

$$Y=rY_r+\left(1-r\right)Y_{1-r} \tag{8}$$

The randomization assumption implies the treatment assignment and the potential outcome variable at baseline are independent. That is, the treatment assignment is balanced with respect to all observed and unobserved baseline confounders. A weaker form of this assumption requires that the potential outcomes be independent of randomization given observed baseline covariates.

In this mediation model, we relax the exclusion restriction and monotonicity assumptions sometimes made under the PS Model (e.g., [22]). The exclusion restriction states the ITT effect is zero for different combination of always mediating and never mediating strata. Hence, no effect of treatment when mediating behavior is constant. Without the exclusion restriction, we can estimate direct effects as ITT effects in the never and always mediating strata. Monotoncity rules out the existence of the defiant mediating stratum. Departures from monotinicty depends on whether there is defiant behavior in the sample. In our mediation context, there are defier individuals, who when assigned to the control group would always not exhibit the positive mediator behavior but when assigned to treatment would always exhibit the positive mediator behavior. For example, in the suicide therapy sample of 101 suicide patients, defiers would be people when assigned to TAU, the control treatment, would always seek outside therapy, and when assigned to the CBT group, would always not seek outside therapy. These people would be observed as defiant mediators. This could be evident for patients in TAU, who have been involved in previous treatments, quite similar to the current TAU, and feel they are not getting substantially better and seek outside therapy to aid their recovery, whereas, for patients in CBT, the treatment may be entirely new, where they feel the treatment is appropriately delivered.

Relaxing this monotonicity assumption increases the dependence of our modeling approach on the model assumptions and covariate-mediation behavior relationships (e.g., [23]). One model assumption to which the PS approach may be sensitive is the constant variance assumption, which we assess with the heterogeneous variance model. In addition, the PS approach depends on relationships involving baseline covariates. Specifically, it assumes no interaction between the baseline covariates, **X**, and the treatment assignment, R, within each principal stratum. Finally, identifiability of the ITT effects within the principal strata is aided with covariates, **X**, that predict and distinguish among these principal strata (e.g., [23]).

2.5 Estimation

Conditional on principal stratum under the above assumptions, the potential outcome models in (1) are equivalent to the following model for the observed outcome Y:

$$Y|C=t,R=r,x\tilde{N}\left(\theta_{ITT_t}r+\gamma^T x,\sigma_t^2\right) \tag{9}$$

We set $\varphi_{ITTt}r+\gamma_t{}^T X = \mu_{tr}$, and define $\boldsymbol{\beta}_\mathrm{t} = [\beta_{t0}\beta_{t1}\gamma^\mathrm{T}]$, therefore, we have $Y|C=t,R=r,\tilde{\mathbf{x}}N\left(\mu_{tr},\sigma_t^2\right)$. We let $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$ represent the vector of principal stratum-specific variances and the principal strata probabilities. We let $\boldsymbol{\beta}$ represent the matrix where the t^{th} row corresponds to the $\boldsymbol{\beta_t}$ for the t^{th} principal stratum. The first two columns correspond to the intercept under the control and treatment, respectively. Hence the difference of these two columns corresponds to θ_{ITTt}, the ITT effect. The remaining columns correspond to the regression coefficients for the covariates within each principal strata identified by the corresponding row.

Estimation for the PS model is based on a mixture of distributions across principal strata.

For subjects assigned to R =0 with mediator D =0 then

$$f\left(Y|D=0,R=0,\pi,\beta,\sigma\right)=\frac{\pi_1}{\pi_1+\pi_3}\phi\left(y|\mu_{10},\sigma_1^2\right)+\frac{\pi_3}{\pi_1+\pi_3}\phi\left(y|\mu_{30},\sigma_3^2\right) \tag{10}$$

which is a mixture across Compliers and Never-Mediators. The symbol ϕ is a normal probability distribution function. For subjects assigned to *R* =1 with mediator *D* =1, then

$$f\left(Y|D=1,R=1,\pi,\beta,\sigma\right)=\frac{\pi_1}{\pi_1+\pi_2}\phi\left(y|\mu_{11},\sigma_1^2\right)+\frac{\pi_2}{\pi_1+\pi_2}\phi\left(y|\mu_{11},\sigma_1^2\right) \tag{11}$$

which is a mixture across Compliers and Always-Mediators. For subjects assigned to *R* =0 with mediator *D* =1, then

$$f\left(Y|D=1,R=0,\pi,\beta,\sigma\right)=\frac{\pi_2}{\pi_2+\pi_4}\phi\left(y|\mu_{20},\sigma_2^2\right)+\frac{\pi_4}{\pi_2+\pi_4}\phi\left(y|\mu_{40},\sigma_4^2\right), \tag{12}$$

which is a mixture across Always-Mediators and Defiers. For subjects assigned to *R* =1 with mediator *D* =0, then

$$f\left(Y|D=0,R=1,\pi,\beta,\sigma\right)=\frac{\pi_3}{\pi_3+\pi_4}\phi\left(y|\mu_{31},\sigma_3^2\right)+\frac{\pi_4}{\pi_3+\pi_4}\phi\left(y|\mu_{41},\sigma_4^2\right), \tag{13}$$

which is a mixture across Never-Mediators and Defiers. Then the posterior distribution is as follows:

$$\begin{aligned}&f\left(\beta,\sigma|R,D,Y,\mathbf{X}\right)\propto p\left(\beta,\sigma,\pi\right)\times\left(\prod_{i:z_i=0,d_i=0}\tfrac{\pi_1}{\pi_1+\pi_3}\phi\left(y|\mu_{10},\sigma_1^2\right)+\tfrac{\pi_3}{\pi_1+\pi_3}\phi\left(y|\mu_{30},\sigma_3^2\right)\right)\times\\&\left(\prod_{i:z_i=1,d_i=1}\tfrac{\pi_1}{\pi_1+\pi_2}\phi\left(y|\mu_{11},\sigma_1^2\right)+\tfrac{\pi_2}{\pi_1+\pi_2}\phi\left(y|\mu_{21},\sigma_1^2\right)\right)\times\\&\left(\prod_{i:z_i=0,d_i=1}\tfrac{\pi_2}{\pi_2+\pi_4}\phi\left(y|\mu_{20},\sigma_2^2\right)+\tfrac{\pi_4}{\pi_2+\pi_4}\phi\left(y|\mu_{40},\sigma_4^2\right)\right)\times\\&\left(\prod_{i:z_i=1,d_i=0}\tfrac{\pi_3}{\pi_3+\pi_4}\phi\left(y|\mu_{31},\sigma_3^2\right)+\tfrac{\pi_4}{\pi_3+\pi_4}\phi\left(y|\mu_{41},\sigma_4^2\right)\right)\end{aligned} \tag{14}$$

where ϕ is a normal distribution parameterized by the specific conditional mean μ_{tz} and variance ${\sigma^2}_t$.

We will employ Bayesian techniques with relatively flat prior distributions to estimate the effects of interest. Markov Chain Monte Carlo (MCMC) techniques are used to implement these Bayesian mixture estimation (e.g., [11,24]). In the case where no covariates are used in modeling the principal strata probabilities, as illustrated in equation (5), we use a Gibbs sampler [25-26] to construct the Markov Chain. The first 100 draws are considered a burn-in. Convergence will be assessed using the Gelman-Rubin method [27]. In the case where covariates are used in modeling the principal strata probabilities, as illustrated in equation (4), we use the Metropolis-Hastings algorithm [28-29] to construct the Markov Chains.

Under the PS model, we specify relatively flat normal and inverse gamma prior distributions for the mean and variance parameters that govern the potential outcomes and principal strata probabilities. First, we assume, $\beta_t \overset{ind}{\square} N\left(\widehat{\beta},\Sigma\right)$, $\sigma_t^2 \square$ *inverse-gamma*(.01,.01). The vector $\widehat{\beta}$ is the estimate from the linear regression of Y on R and **X**, and $\mathbf{\Sigma} = \mathbf{n}\mathbf{V}_{\beta}$ where $\mathbf{V}_{\beta}$ is the variance-covariance matrix of $\widehat{\beta}$. The data-based priors for θ_{ITTt} follow similar strategies by Hirano et al. [24] and Ten Have et al. [12]. Different prior distributions were specified for the logistic models for the principal strata depending on whether covariates were used to help predict the principal strata membership. With covariates, we again specified a relatively flat multivariate normal distribution under different prior variances specified for the intercept and each covariate in the prediction of principal strata membership as illustrated in equation (4). The different prior variances for the coefficients in equation (4) account for the different scaling for the corresponding covariate [24]. In the absence of covariates predicting the principal strata, the relatively flat Dirichlet prior is used. The distributions from which parameters are drawn at each iteration of MCMC are illustrated in the appendix.

2.6 Software for estimating the PS model

As discussed by Bellamy et al. [30], a SAS macro for the PS approach is available at http://www.cceb.upenn.edu/pages/tenhave/files.html within Causal SAS macros with documentation in the README.TXT text file within the zip file. This software has been used in the published analyses in Ten Have et al. [12], Bellamy et al. [30], and Lynch et al. [31].

3. Simulations

To assess the accuracy and robustness of our estimation procedures, we present four sets of simulations under the principal stratification model. All simulated data sets were based on the data from the sample of 101 adults who attempted suicide and were randomized to CBT or usual care. The first two sets of simulations were based on the parameter estimates from fitting the principal stratification model with a separate variance for the defiant mediating principal stratum (heterogeneous variance by defier model), where the first and second sets corresponded to the original sample size of 101 and then a doubling of the sampling size to 202, respectively. The third and fourth sets of simulations were based on the same analysis model but which was mis-specified for the true model for which a separate variance was specified for the always mediating principal stratum instead of the defiant mediating class (heterogeneous variance by always mediator model). The sample sizes of the third and fourth sets paralleled the sample sizes of the first and second set. The true simulation models for all of these simulations were based on the following specifications from the analysis of the actual data:

- Principal strata probabilities of 0.024 for compliant mediators; 0.129 for always mediators; 0.752 for never mediators; and 0.095 for defiant mediators;
- Bernoulli treatment assignment randomly generated with probability of a CBT versus usual care designation of 0.50;
- Baseline depression randomly generated as a univariate normal variable with mean of 31.9 and standard deviation of 13.8;
- Endpoint depression generated as a univariate normal variable with mean dependent on treatment assignment, principal stratum, and a baseline covariate with $\beta = \begin{bmatrix} 4.62 & -2.91 & 0.50 \\ 11.21 & 1.10 & 0.35 \\ 3.38 & -1.62 & 0.55 \\ 3.03 & 0.08 & 0.001 \end{bmatrix}$ and a pooled standard deviation of 12.0 for the compliant, always, and never mediating strata and a standard deviation of 0.8 for the defiant mediating stratum (Rows for **β** correspond to each principal strata ordered in accordance to equation (1). The first two columns correspond to the intercept under control, and treatment, respectively. The third column corresponds to the baseline covariates regression coefficient within each stratum, respectively).

For the third and fourth simulations based on the mis-specified model, we used the same above specifications, except we specified a standard deviation structure that was different for the true and fitted models. The true model assumed a standard deviation of 12.0 for the compliant mediating, never mediating, and defiant mediating strata and a standard deviation of 0.8 for the always mediating stratum.

With the above specifications, simulated depression outcomes, baseline depression status, and treatment and principal strata assignments were generated. For each of 500 generated data sets, we fitted the model with 10,000 MCMC chains after 100 burn-in iterations. We fitted both the heterogeneous variance by defiers and homogeneous variance principal stratification models. Simulations are summarized in Table 1. The summary measures for the direct effect parameter are presented separately for the never and always mediating classes as well as for a weighted average of these two classes:

- the coverage of nominal 95% confidence interval, which is the percentage of iterations for which the corresponding nominal 95% confidence interval includes the true parameter;
- the average bias, which is difference between average estimate of the effect with the true value, -5.70 for the direct effect, -10.05 for the always mediator ITT effect, and -4.96 for the never mediator ITT effect;
- the percent bias, which is the magnitude of the difference between the true parameter value and average estimate divided by the true value, then multiplied by 100%; and
- the mean squared error (MSE) for average estimate.

Table 1 displays results for the four simulations, where the upper portion corresponds to the correctly specified analysis model and the lower portion corresponds to the mis-specified analysis model. For each of the upper and lower portions, the first row corresponds to the results for simulations based on a sample size of 101, and the second row corresponds to the results based on a sample size of 202.

The correctly specified heterogeneous variance by defiers model does produce stable estimates of the direct effects in the always and never taker strata, with coverage in line with the nominal 95% level. For the never mediating stratum, coverage is 95.0% and 95.3% for the simulations based on sample sizes of 101 and 202 respectively. Similarly, for the always mediating stratum, coverage is 97.6% and 97.9%. We also see the stable estimates for the overall direct effect, the weighted average of these two strata, which has coverage of 96.2% and 95.3% for the simulations based on sample sizes of 101 and 202, respectively.

Comparison of the heterogeneous variance by defiers model with the homogeneous variance model indicates a better fit for the heterogeneous variance model as illustrated by less bias in the corresponding effects, although coverage is quite comparable between the two models. For the never mediating stratum based on the 101 simulation sample size, percent bias diminished from 32.5% to 15.8%, while for the always mediator class, percent bias decreased from 50.1% to 18.7%. For the overall direct effect, percent bias decreased from 10.4% to 6.6%. A similar pattern holds true for simulations based on sample size of 202.

To determine the robustness of inference with respect to mis-specification of the heterogeneous variance analysis model, we fitted the homogeneous variance and heterogeneous variance by defiers models to data generated from a heterogeneous variance model with a different variance specified for the always mediating class rather than the defiant mediator stratum. The results revealed an increase in bias for the effects estimated under the heterogeneous variance analysis model. An increase in bias is also seen for the effects estimated under the homogeneous variance model with the exception of the ITT effect in the never mediating stratum. Within this class, we do see an increase in the MSE for the mis-specified heterogeneous variance by defiers model compared to the correctly specified heterogeneous variance by defiers model. Coverage was worse for both the homogeneous and heterogeneous variance by defiers models when the true model assumed a different variance for the always mediators. In particular, for the always mediating stratum based on the original overall sample size of 101, we see coverage was reduced by more than 50% for the heterogeneous variance by always mediator model and by 20% for the homogeneous variance model. There was less of reduction in the coverage for the never mediating class, for which the class probabilities were significantly larger than the probabilities for the always mediators. This result suggests that the negative impact of model mis-specification may be most significant for strata with such small membership probabilities.

4. Applications

We illustrate the mediation process on the two examples discussed earlier. We present PS model-based estimates of the direct effects in the always and never mediator principal strata for the two example datasets with a comparison to a standard regression approach to estimating direct effects for the whole sample under the assumption of sequential ignorability. These direct effect estimates are compared in the context of the overall ITT estimate for the whole sample using standard regression with the outcome variable as the dependent variable and randomized indicator variable as the primary covariate, while adjusting for the baseline outcome. We also assess the sensitivity of the PS approach to the homogeneous variance assumption.

4.1 Cognitive therapy treatment for suicide attempters

A goal for CBT is for patients to have a better understanding of their depression and mechanisms of an onset of a depression episode and also develop cognitive strategies for dealing with these mechanisms rather than reliance on other types of strategies such as psychotherapy. Hence, CBT patients should be less likely to seek external therapy dealing with psychology of the underlying problems rather than cognitive issues. Accordingly, the CBT effect compared to TAU effect is working through less reliance on outside therapy such as psychotherapy

To assess the direct effect of CBT apart from reducing reliance on external therapy, we proceed with the methods discussed in section 2. Baseline BDI was adjusted for as a covariate in all models. Using the notation described in the previous section we have the following:

- R = 1 for CBT and R = 0 for TAU
- D = 1 for those who receive outside therapy, D = 0 for those who do not receive outside therapy.
- Y is the BDI score at 6 months.
- X, the vector of baseline covariates, consists of the baseline BDI score.

The Bayesian estimation process was implemented based on flat prior distributions described in Table 2. Baseline covariates were adjusted for in the outcome models, but not the prediction model for the principal strata. The data analysis results were very similar between adjusting and not adjusting for covariates in the principal strata model. With no adjustment for covariates in the model for the principal strata in equation (5), the homogeneous and heterogeneous variance PS models were estimated with MCMC as implemented through a Gibbs sampler. For each model, the first 100 draws were considered a burn-in. We used 45,000 iterations with convergence shown with the Gelman-Rubin statistic. Posterior distributions were generated for all effects. Posterior confidence intervals for each ITT effect within principal stratum were generated. Direct effects for treatment within the never and always mediator principal strata were derived based on the posterior distributions for the effects in equation (3). Under both the heterogeneous and homogeneous variance PS models, we estimated the pooled direct effect of treatment across the always and never mediator strata based on equation (7).

For the heterogeneous variance model, where variances varied across principal strata, we found that for all principal strata except the defier-mediator principal stratum, small sample sizes of 10 or less occurred occasionally across the Gibbs Sampling iterations, given the probabilities for these principal strata were consistently low. Consequently, we considered a heterogeneous variance by defiers model, under which the variance for the defiant mediating stratum differed from the variance for the other three principal strata, while the other three principal strata have one common variance. The choice of the defiant mediating stratum is also based on our relaxation of the monotonicity assumption. While our samples do exhibit defiant mediating behavior, clinically, it would seem these people are somewhat different compared to those who are compliant mediating, never mediating, or always mediating. We assess this assumption by comparing the corresponding results to those under the PS with homogeneous variances and under the standard regression approach of Baron and Kenny [4], which estimates direct effects for the whole sample. The results are illustrated in Table 3:

Table 3 shows that the standard regression direct effect for the overall sample is nearly comparable to the ITT treatment effect with 95% confidence intervals of (-12.01,-1.70) and (-11.37,-1.73), respectively. Hence, it appears that under the standard approach, most of the intent-to-treat effect of the CBT intervention is not through the mediator, outside therapy.

Under the homogeneous variance PS model, we see that the direct effect estimates in both the always and never mediating principal strata, and consequently the pooled direct effect estimate, exceed the overall ITT effect. The wide confidence intervals for the PS stratum-specific direct effects notwithstanding, the ITT effects in the non-informative principal strata (complier and defier mediators) are smaller than the never and always mediator direct effects. However, the PS-based wide confidence intervals for the always and never mediating classes overlap with the standard ITT and direct effect estimates, as does the pooled PS confidence interval. The pooled PS ITT confidence interval is narrower as one would expect with a larger sample size, but still does not quite show statistical significant, as the nominal 95% confidence interval is (-13.96, 0.01),

In contrast to the large pooled direct effect estimate under the homogeneous variance model, the heterogeneous variance by defier PS model yielded a reduced pooled direct effect estimate relative to the overall ITT estimate and the standard direct effect for the whole sample. The heterogeneous variance PS direct effect estimate was -5.70, 20% less than the homogeneous variance direct effect estimate of -7.11 However, the 95% confidence interval of (-10.92, -0.49) for the heterogeneous variance pooled direct effect was narrower and therefore significant relative to the marginally non-significant analogous homogeneous variance interval of (-13.96, 0.01). In addition, the pooled direct effect estimate under the homogeneous variance model was more than 15% less than the overall ITT estimate and the standard regression direct effect estimate. The smaller heterogeneous variance PS direct effect estimate was due to a smaller direct effect estimate in the heavily weighted never mediator stratum under this PS model than under the homogenous variance model. The relatively small pooled direct effect estimate (-5.70) under the heterogeneous variance model is still larger than in magnitude than the rank preserving model-based estimate (-3.93) for the whole sample reported by Ten Have et al. [6]. This difference may be due to the potential heterogeneity of direct effects in the sample as represented by the differences between always and never mediators addressed next.

Compared to the corresponding estimates under the homogeneous variance model, the heterogeneous variance PS model exhibited more heterogeneity between the always and never mediator strata-specific direct effect estimates in Table 3, although the confidence intervals still overlap because of their wide widths. The confidence intervals notwithstanding, the stark difference between the ITT effect estimates in the always and never mediator strata suggest the possibility of a CBT intervention-outside therapy interaction. That is, the CBT intervention may be more effective in the always mediating group than in the never mediating group. We expect CBT patients to be less likely to seek additional therapy due to CBT focusing on an individual's understanding of the mechanism and coping strategies for a depression onset. However, for those who are very committed to seeking outside therapy (always mediating group), you would expect them to be more receptive to any type of therapy and therefore CBT should be more effective in this more committed group. For the group that would never seek outside therapy, you would expect them to also be less receptive to CBT and therefore CBT should less effective in this group.

Given the difference in direct effect estimates between the heterogeneous and homogeneous variance PS models, the fit of the models have been compared as follows. To assess convergence of the MCMC chains under the homogeneous and heterogeneous variance models, we employed the Gelman-Rubin Statistic For the homogenous model the GR statistic ranged from 1.00003 to 1.0024 for various sets of starting values of the chain. Similarly, for the heterogeneous variance model the GR statistic ranged from 1.00048 - 1.0165.

Figures 3 and 4 present plots of the prior and posterior distribution of θ_{ITTt} for each principal stratum. According to criteria formulated by Garret and Zeger [32] for latent class models such as the PS model, the substantial discrepancies between posterior and prior distributions for the principal strata indicate the data contributed information to identification of the ITT estimates for the specific principal strata. Figure 3 plots the distributions for the homogeneous model and Figure 4 plots the distribution for the heterogeneous variance model.

As seen in Figure 3 for both the compliant and always mediating strata, the posterior is close to the prior distribution, suggesting that the data did not provide much information for identifying the ITT effect in these principal strata. However, for the never and defiant mediating strata, there is sufficient separation between the prior and posterior distribution, suggesting that the data provided information for identifying the ITT effect in these two strata. In Figure 4, for the compliant mediators, the posterior is close to the prior distribution, suggesting that the data did not provide much information for identifying the ITT effect in that group. In contrast, for the always, never, and defiant mediating strata, there is sufficient separation between the prior and posterior distribution, suggesting that the data provided information for identifying the ITT effect in these three stratum. These results may suggest that the heterogeneous variance model identifies the data better than the homogeneous variance model.

4.2 Collaborative Care suicide prevention study

In the Bruce et al. [16] study, the practices with a Depression Specialist (DS) may be more likely to recognize an onset of a depression episode, resulting in immediate prescription of medication compared to the practices without a DS. This could lead to better depression status for patients randomized to practices with a DS compared to patients randomized to practices without a DS. However, the presence of a DS in practice may have a direct effect on reducing depression symptoms even for patients who would always take medication regardless of the presence of a DS (always-medicators) or never take medication (never-medicators). This direct effect would occur if the presence of the DS in a practice were to increase overall efforts by practice providers and staff in treating depression using other strategies in addition to medication.

To assess the direct effect of the DS apart from increasing medication use, we proceed with the methods discussed in section 2. Because the within-practice design effect is so small for the outcome, we ignored the clustering due to primary practice as was done in previous publications of the study results [16,33]. Using the notation described in the previous section we have the following:

- R = 1 for patients randomized to a practice with a depression specialist and R = 0 for patients randomized to a practice without a depression specialist
- D = 1 for those who took medication, D = 0 for those who did not take medication.
- Y is the HRSD score at 4 months.
- **X**, the vector of baseline covariates, consists of the baseline HRSD score, and baseline suicide ideation.

The standard ITT analysis for the whole sample of the outcome HRSD at 4 months while controlling for baseline HRSD and baseline suicide ideation [16] is significant. To asses whether this significant ITT effect due to the presence of the DS in the primary care practice occurred around the mediator, whether patients were taking medication, we compare the standard direct effect estimate for the overall sample under the standard regression approach of Baron and Kenny [4] with various PS model-based estimates for always and never mediators with and without homogeneous variance assumptions. Unlike the previous example, the sample size of this example is sufficient to estimate a separate variance for each principal stratum (heterogeneous variance model), so no pooling variance across select principal strata is necessary.

The Bayesian estimation process for the PS models was implemented based on flat prior distributions as described in Table 4, using the Metropolis-Hastings algorithm to adjust for covariates under the principal strata multinomial model in equation (4) The first 100 draws were considered a burn-in. Again, we used 49,000 iterations with convergence shown with the Gelman-Rubin statistic. Posterior distributions intervals were generated for all effects. Posterior means and confidence intervals for each ITT effect within each principal stratum generated. The pooled direct effect for treatment across never and always Mediators was derived based on the posterior distributions for the pooled effect in equation (7). The results for the overall ITT effect, the standard regression overall direct effect, and stratum-specific and pooled estimates under both PS models are illustrated in Table 5:

The estimated direct effects of the DS intervention under the different modeling approaches differ more in terms of inference than in terms of magnitude. While the standard regression direct effect estimate is approximately 15% less than the overall ITT effect estimate, suggesting some mediation, the direct effect estimate is still significant. In contrast, the pooled direct effect estimates for the always and never mediator strata under both the homogeneous and heterogeneous variance PS models are larger than the overall ITT effect estimate, although their confidence intervals surround zero. The larger PS direct effect estimates are inflated because of the very large direct effect estimates in the never mediator (never medicated) stratum under both variance PS models. With such a small percentage (3.9% per Table 5) of the sample potentially belonging to the never medicated group, one may want to base inference on the always medicated stratum under either of the variance PS models. With non-significant confidence intervals, the direct effect estimates under the two models range between -2.83 and -2.88, about 10% less than the overall ITT estimate of -3.12. The non-significant confidence intervals notwithstanding, the direct effect estimates in the always medicated group are in some agreement with the standard regression direct effect estimate, all being 10-15% less than the overall ITT estimate. This consistency between PS and standard direct effect estimates corresponds to a similar pattern observed for the rank preserving (-2.58) and standard direct effect (-2.67) estimates in Ten Have et al. [6].

Keeping in mind the wide confidence intervals, the stark difference between the ITT effect estimates in the always and never mediator strata under both PS models suggest the possibility of a DS intervention-medication interaction. That is, the DS intervention may be more effective in the never medication group than in the always medication group. However, under the heterogeneous and homogeneous variances, the 95% confidence intervals of the contrast in the DS effect between the always and never mediation groups contain 0: (-16.83, 23.49) and (-8.95,12.96) for the homogeneous and heterogeneous variance models, respectively. Hence, there is not enough evidence to show an intervention-mediator interaction.

We again used the Gelman-Rubin [27] method to assess the convergence of the MCMC chains. For the homogenous model, the GR statistic ranged from 1.0011-1.0048 for various sets of starting values. Similarly, for the heterogeneous variance model, the GR statistic ranged from 1.0005 to 1.0021.

Figures 5 and 6 offer plots of the prior and posterior distributions of θ_{ITTt} for each principal stratum. Figure 5 plots the distributions for the homogeneous model, and Figure 6 plots the distribution for the heterogeneous variance model.

Both Figures 5 and 6 indicate there is sufficient separation in each of the four principal strata between the prior and posterior distributions. Hence, there is evidence to suggest that both the homogeneous and heterogeneous variance models appear to identify the data well.

5. Discussion

In addressing the direct effect of treatment in the context of mediation, the presented examples and simulation results illustrate that care must be taken in interpreting standard direct effect estimates under the sequential ignorability assumption and complementing these estimates with causal estimates of direct effects that relax such an assumption. In this paper, we compared such causal estimates under the PS approach with the standard regression direct effect estimates in a simple context with a single continuous outcome and binary intervention and mediator. In one example, there was agreement, and in the other there was greater disagreement between the standard and causal approaches. Moreover, we assessed the sensitivity of the PS approach to departures from the homogeneous variance assumption. In the example with agreement between the PS and standard approaches, the PS method appeared to not be sensitive to the variance assumption. In contrast, in the example with a smaller sample, there was disagreement between the homogeneous and heterogeneous variance PS model results.

With the PS model as with all causal approaches, there are limitations, which may have led to differences between the results of the PS and standard regression direct effect approaches. One standard assumption is a common variance in the potential outcome error distribution across principal strata. In this investigation, we relaxed this assumption allowing heterogeneity in the variance of the error distribution across principal strata. Our simulations showed somewhat increased bias and poorer coverage under mis-specification of the principal stratum-specific variance structure. From a modeling perspective investigators may want to explore the relaxation of the homogeneity of variance assumption during their implementation. Comparison of the stability of the effects as well as assessment of rate of convergence and model identifiability under homogeneity and heterogeneity will help the investigator assess the need for modeling heterogeneity.

In addition, the PS approach is dependent on the validity of unverifiable assumptions. SUTVA, may not hold when interventions require each clinician/therapist treating multiple patients as in both studies; therefore, the interchangeability of subjects may be somewhat questionable. The randomization assumption may also be somewhat implausible in the Collaborative Care Study, for which primary care practices were randomized. An imbalance for an observed baseline variable, suicide ideation, occurred and was adjusted for. Given that the randomization units were the 20 primary care practices in the study, such an imbalance is not unlikely. Nonetheless, this was the only one of many baseline covariates that exhibited significant differences between the randomized groups. Hence, there is equivocal evidence of potential unmeasured confounding.

The additional assumption of no interactions involving the randomized intervention and mediator with baseline covariates conditional on the principal stratum is crucial for identifying parameters under the PS approach. In the suicide cognitive therapy study, the cognitive therapist may have provided more intensive therapy for patients with more suicide ideation or depression at baseline, which would have resulted in an R*X interaction. Nonetheless, the investigators believed that the cognitive therapy approach is standardized enough that such an interaction was not very likely. Similarly, baseline depression and suicide ideation may have impacted the way an external therapist provided the mediator, non-study therapy. However, this interaction may also be unlikely given that non-study therapists mostly saw patients after the study started and may not be aware of the patients' baseline values. In the Collaborative Care Study, R*X and D*X interactions may have more likely existed given that the DS intervention was less scripted than the CBT intervention in the other study and the mediator, medication adherence, is a patient factor and thus possibly sensitive to baseline depression and ideation.

A benefit of the PS model is that it does not require the assumption of no structural interaction between the treatment assignment and the mediator. Comparing the direct effects between the always and never mediating principal strata, the two strata for which the potential mediation level is constant, we can assess a component of the treatment assignment by mediator interaction. A substantial difference in the direct effect between these two strata would suggest such an interaction.

This paper focuses on direct effects on a follow-up continuous outcome at a single follow-up visit. The PS approach has been extended to longitudinal data in the context of treatment non-compliance in randomized trials (e.g., [34]). Using such longitudinal extensions in the mediation context is an area of future research. In addition, the PS approach has been extended to binary outcomes in the non-compliance context (e.g., [24]). Similar extensions to binary outcomes in the mediation context are needed.

While the limitations of the PS approach have been addressed above, this approach offers a viable technique for assessing mediation when the mediator is not randomized, i.e. lack of sequential ignorability. The PS approach can provide evidence for mediation but only for the always- and never-mediator classes. With the caveat of potentially low power, no direct effect in the always- and never-mediator classes is indicated when the confidence intervals for the ITT effects in these classes include zero but are narrow enough for precise inference. Assuming that one can manipulate the mediator such that subjects in the always-mediator group can be made to exhibit never-mediator behavior and vice-versa for subjects in the never-mediator group, the absence of a direct effect in these classes implies that the resulting ITT effects within these classes are totally mediated by the mediator. Also, all mediation approaches pertaining to all types of subjects make the assumption of a mediator that can be manipulated [2]. MacKinnon et al. [35] describe causal inferences as new approaches for mediation. In fact, they highlight the PS approach of Frangakis and Rubin [10] as a promising method, as do Lynch et al. [31]. Finally, the PS approach has an additional advantage over more traditional approaches beyond the benefit of protection against unmeasured confounding. The PS approach accommodates mediators that are not under the control of the investigator or clinician. In our example, the traditional approach assumes that the clinician has control over whether the patients seek outside therapy.

Acknowledgements

The preparation of this manuscript was funded in part by R01MH078016.

Appendix

MCMC Draws

We draw C_i from a multinomial distribution where the probability for each class t, is $P\left(C_i=t|R_i=r,D_i=d,Y_i,\beta,\sigma^2\right)=\frac{\pi_i\phi\left(y_i|\mu_{tr},\sigma_t^2\right)}{\pi_t\phi\left(y_i|\mu_{tr},\sigma_t^2\right)+\pi_{t^*}\phi\left(y_i|\mu_{t^*r},\sigma_{t^*}^2\right)}$ where t^* corresponds to the other principal stratum for the pair $R_i = r$, $D_i = d$.

The distributions from which the other parameters drawn at each iteration are as follows:

$$\begin{aligned}&\left(\beta|\mathbf{X},\mathbf{Y},\sigma^2\right)\sim MVN\left(\widehat{\beta},\Sigma\right)\\&\widehat{\beta}=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{Y}\\&\widehat{\Sigma}=\sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\end{aligned}\tag{A1}$$

$$\begin{aligned}&\left(\sigma^2|\mathbf{X},\mathbf{Y},\beta\right)\sim INV-\chi^2\left(df=0.01+N,\tfrac{0.01^2+v}{0.01+N}\right)\\&v=(\mathbf{Y}-\mathbf{X}\beta)^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\beta)\end{aligned}\tag{A2}$$

$$\begin{aligned}&\left(\pi_t|\mathbf{X},\mathbf{Y},\mu,\sigma^2\right)\sim\text{Dirichlet}\,(n_t+1)\\&n_i=\sum_{i=1}^{N}I\,(C_i=t)\end{aligned}\tag{A3}$$

where N is the sample size and I is the indicator function which equal 1 when the equation within the parentheses is true and 0 otherwise, the matrix **X** includes baseline covariates augmented with indicators for control and treatment assignment, and **Y** is the vector of outcomes.

References

1. Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics 2002;58:21–29. [PubMed: 11890317]

2. Rubin D. Direct and indirect causal effects via potential outcomes. Scandinavian Journal of Statistics 2004;31:161–170.

3. Mealli F, Imbens GW, Ferro S, Biggeri A. Analyzing a Randomized Trial on Breast Self-Examination with Noncompliance and Missing Outcomes. Biostatistics 2004;5:207–222. [PubMed: 15054026]

4. Baron RM, Kenny DA. The moderator-mediator distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 1986;51:1173–1182. [PubMed: 3806354]

5. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychological Methods 2002;7:83–104. [PubMed: 11928892]

6. Ten Have T, Joffe M, Lynch K, Maisto S, Brown G, Beck A. Causal mediation analyses with rank preserving models. Biometrics 2007;63:926–934. [PubMed: 17825022]

7. Robins J, Rotnitzky A. Estimation of treatment effects in randomized trials with non-compliance and dichotomous outcome using structural mean models. Biometrika 2005;91:763–783.

8. Imbens G, Rubin D. Bayesian inference for causal effects in randomized experiments with noncompliance. Annals of Statistics 1997;25:305–327.

9. Jo, B.; Muthen, B. Longitudinal Studies with Intervention and Noncompliance: Estimation of Causal Effects in Growth Curve Mixture Modeling. In: Duan, N.; Reise, SP., editors. Multilevel Modeling: Methodological Advances, Issues, and Applications. Lawrence Erlbaum; New York: 2001. p. 51-62.

10. Frangakis CE, Brookmeyer RS, Varadhan R, Safaeian M, Vlahov D, Strathdee SA. Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a Needle exchange Program. Journal of the American Statistical Association 2004;97:284–292.

11. Mealli F, Rubin D. Commentary: Assumptions allowing the estimation of direct causal effects. Journal of Econometrics 2003;112:79–87.

12. Ten Have T, Elliot MR, Joffe M, Zanutto E, Datto C. Causal models for randomized physician encouragement trials in treating primary care depression. Journal of the American Statistical Association 2004;99:16–25.

13. Dunn G, Bentall R. Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). Statistics in Medicine. in press

14. Brown G, Ten Have T, Henriques G, Xie SX, Hollander EJ, Beck AT. Cognitive Therapy for the Prevention of Suicide Attempts: A Randomized Controlled Trial. Journal of the American Medical Association 2005;294:2847–2848.

15. Beck, AT.; Steer, RA.; Brown, GK. Manual for the BDI-II. The Psychological Corporation; San Antonio, TX: 1996.

16. Bruce ML, Ten Have TR, Reynolds CF, Katz II, Schulberg HC, Mulsant BH, Brown GK, McAvay GJ, Pearson JL, Alexopoulos GS. A Randomized Trial to Reduce Suicidal Ideation and Depressive Symptoms in Depressed Older Primary Care Patients: The PROSPECT Study. Journal of the American Medical Association 2004;291:1081–1091. [PubMed: 14996777]

17. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56–62. [PubMed: 14399272]1960

18. Williams JB. A structured interview guide for the Hamilton Depression Rating Scale. Arch Gen Psychiatry 1988;45:742–747. [PubMed: 3395203]

19. Neyman J. On the application of probability theory to agricultural experiments: essay on principles, section 9. Statistical Science 1990;5:165–480.translated in

20. Rubin DB. Bayesian inference for causal effects - Role of randomization. Annals of Statistics 1978;6:34–58.

21. Rubin D. Statistics and Causal Inference: Comment Which Ifs Have Causal Answers. Journal of the American Statistical Association 1986;81:961–962.

22. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 1996;91:444–455.

23. Rubin D. Causal inference through potential outcomes and principal stratification. Statistical Science 2007;21:299–309.

24. Hirano K, Imbens GW, Rubin DB, Zhou XH. Assessing the effects of an influenza vaccine in an encouragement design. Biostatistics 2000;1:69–88. [PubMed: 12933526]

25. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian restoration of imagages. IEEE Transactions on pattern analysis and machine intelligence 1984;6:721–741.

26. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 1990;85:398–409.

27. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 1992;7:457–472.

28. Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 1970;57:97–109.

29. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. Vol. 2nd ed.. Chapman and Hall; New York: 2004.

30. Bellamy SL, Lin JY, Ten Have TR. An introduction to causal modeling in clinical trials. Clinical Trials 2007;4:58–73. [PubMed: 17327246]

31. Lynch KG, Cary M, Gallop R, Ten Have TR. Causal Mediation Analyses for Randomized Trials. Health Services and Outcome Methodology 2008;8:57–76.

32. Garrett ES, Zeger SL. Latent Class Model Diagnosis. Biometrics 2000;56:1055–1067. [PubMed: 11129461]

33. Small D, Ten Have T, Joffe M, Cheng J. Random effects models for analysing efficacy of a longitudinal randomized treatment with non-adherence. Statistics in Medicine 2006;25:1981–2007. [PubMed: 16220487]

34. Lin J, Ten Have T, Elliott M. Longitudinal Nested Compliance Class Model in the Presence of Time-Varying Noncompliance. Journal of the American Statistical Association 2008;103:462–473.

35. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation Analysis. Annual Review of Psychology 2007;58:593–614.

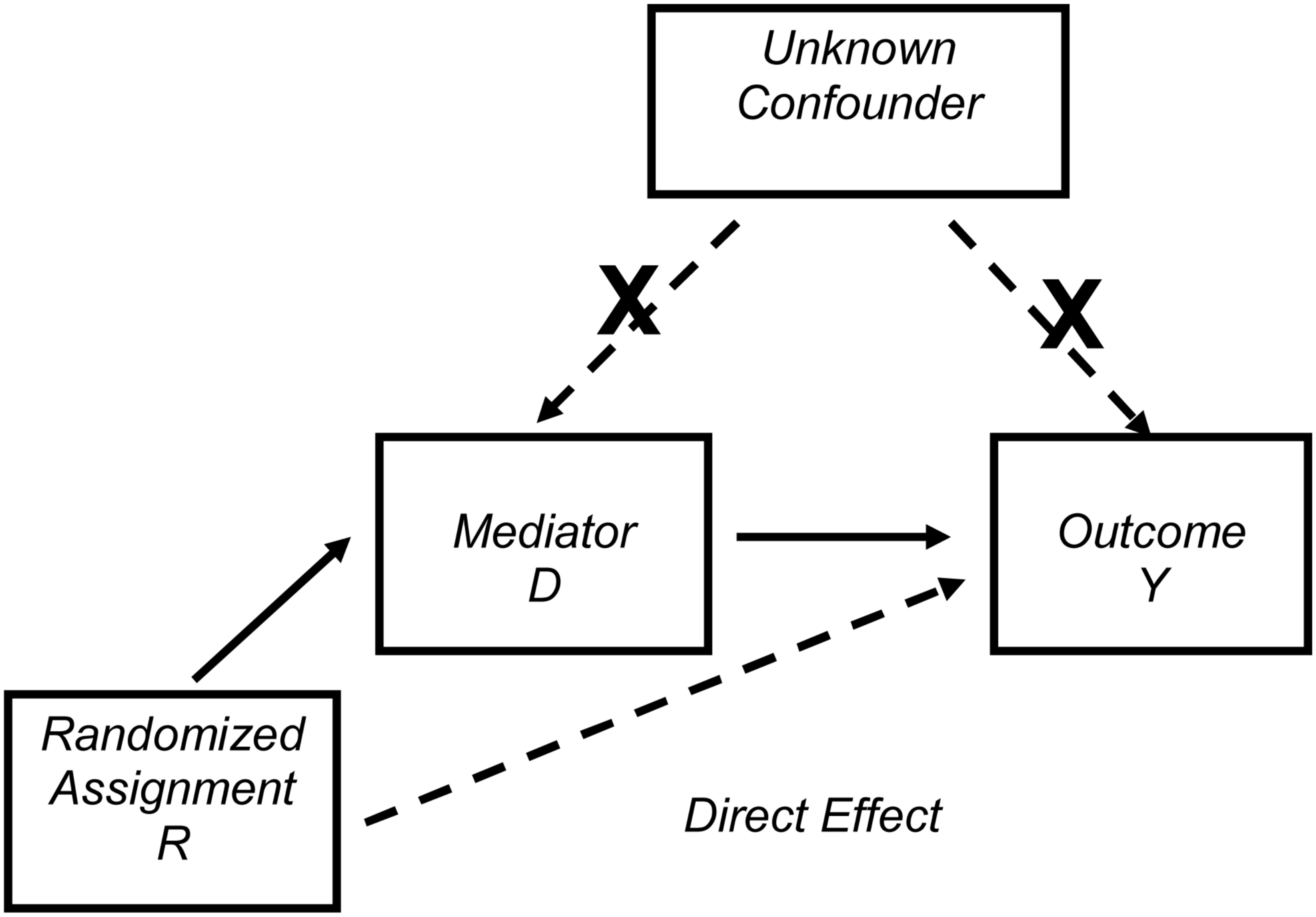

Figure 1. Mediation Process

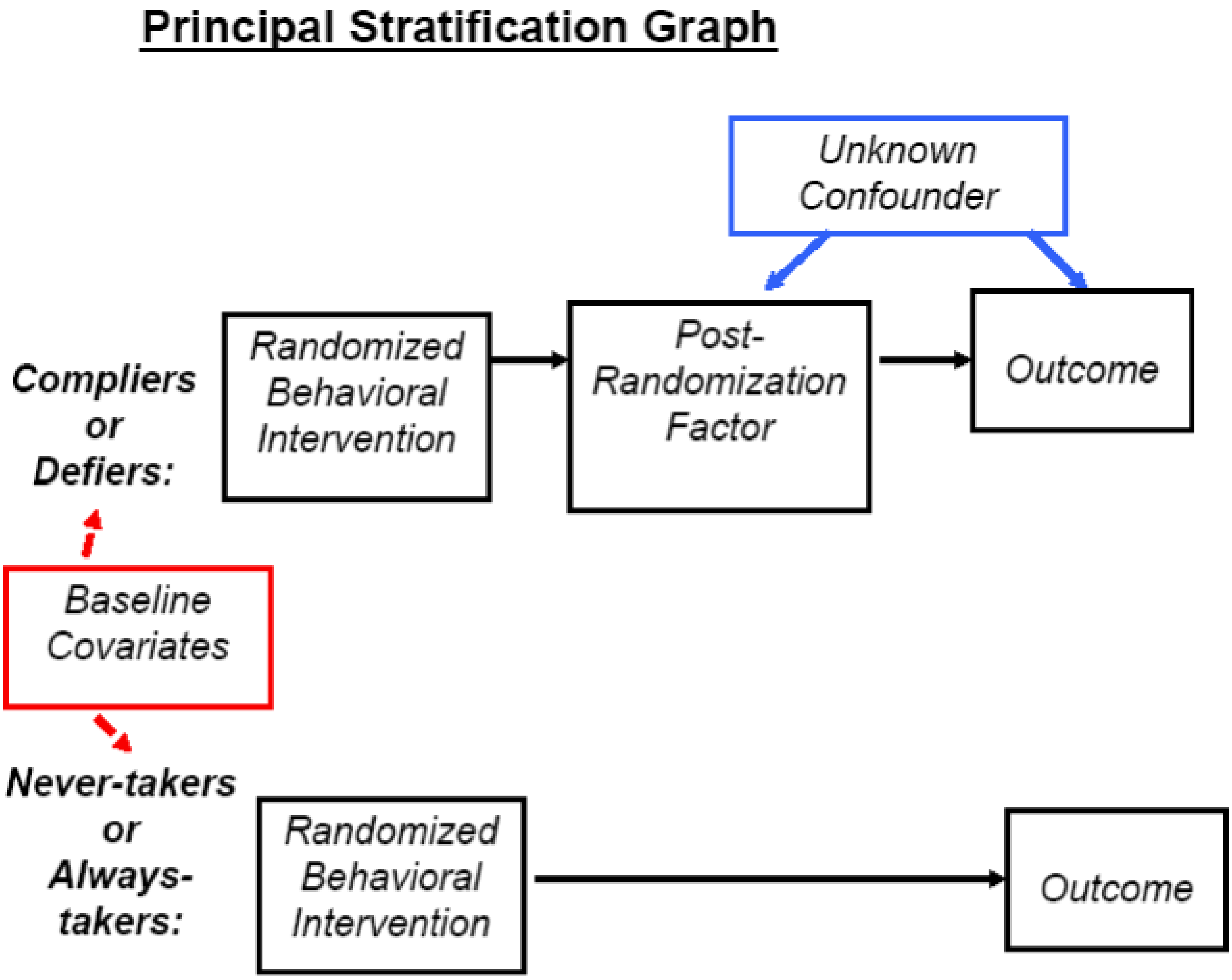
Principal Stratification Graph
Unknown Confounder
Compliers or Defiers:
Randomized Behavioral Intervention
Post-Randomization Factor
Outcome
Baseline Covariates
Never-takers or Always-takers:
Randomized Behavioral Intervention
Outcome

Figure 2.
Mediation process with the four latent classes of the Principal Stratification Model

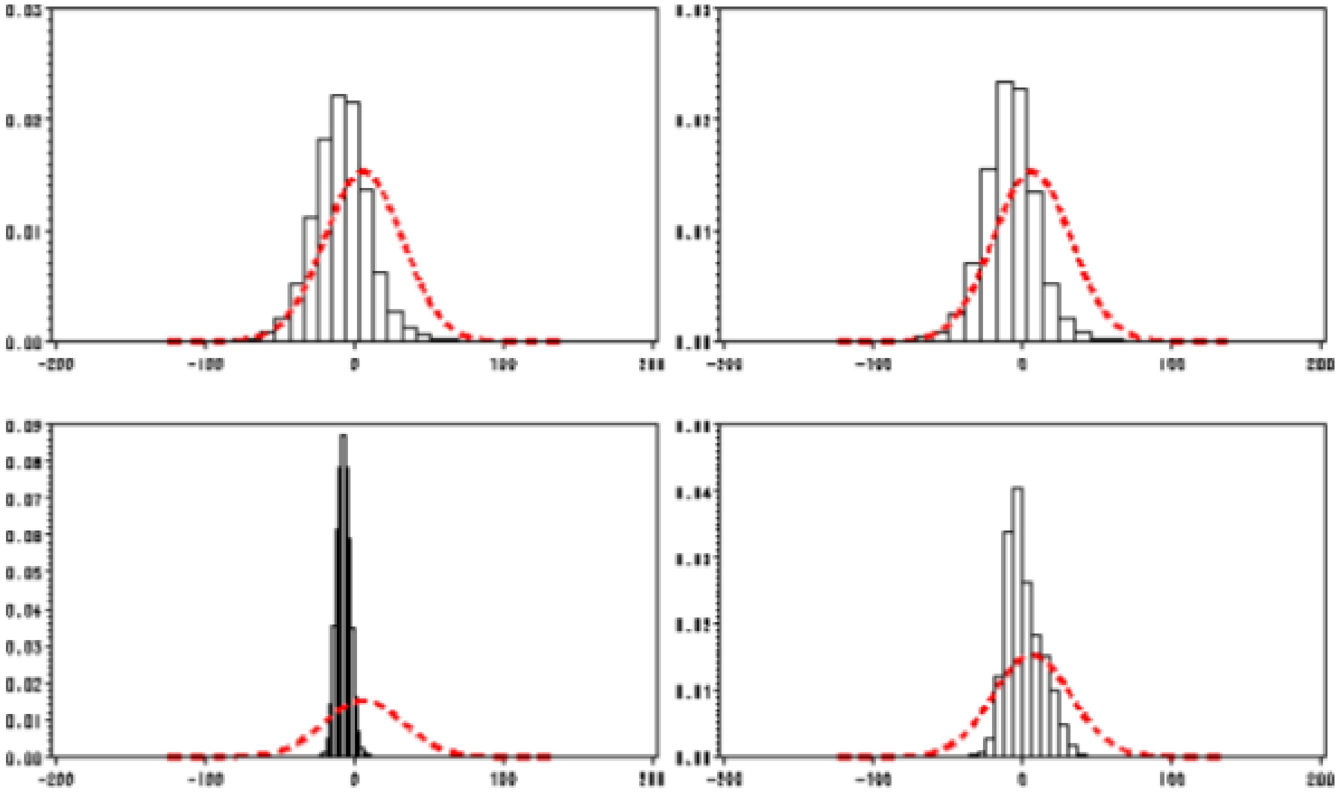

Figure 3.

Plots of the Prior (smooth line) and Posterior (histogram) distribution of the ITT estimates for each principal strata for the homogeneous variance model.

Note: Plots from left to right and top to bottom are: Compliant mediating, Always mediating, Never mediating, and Defiant mediating.

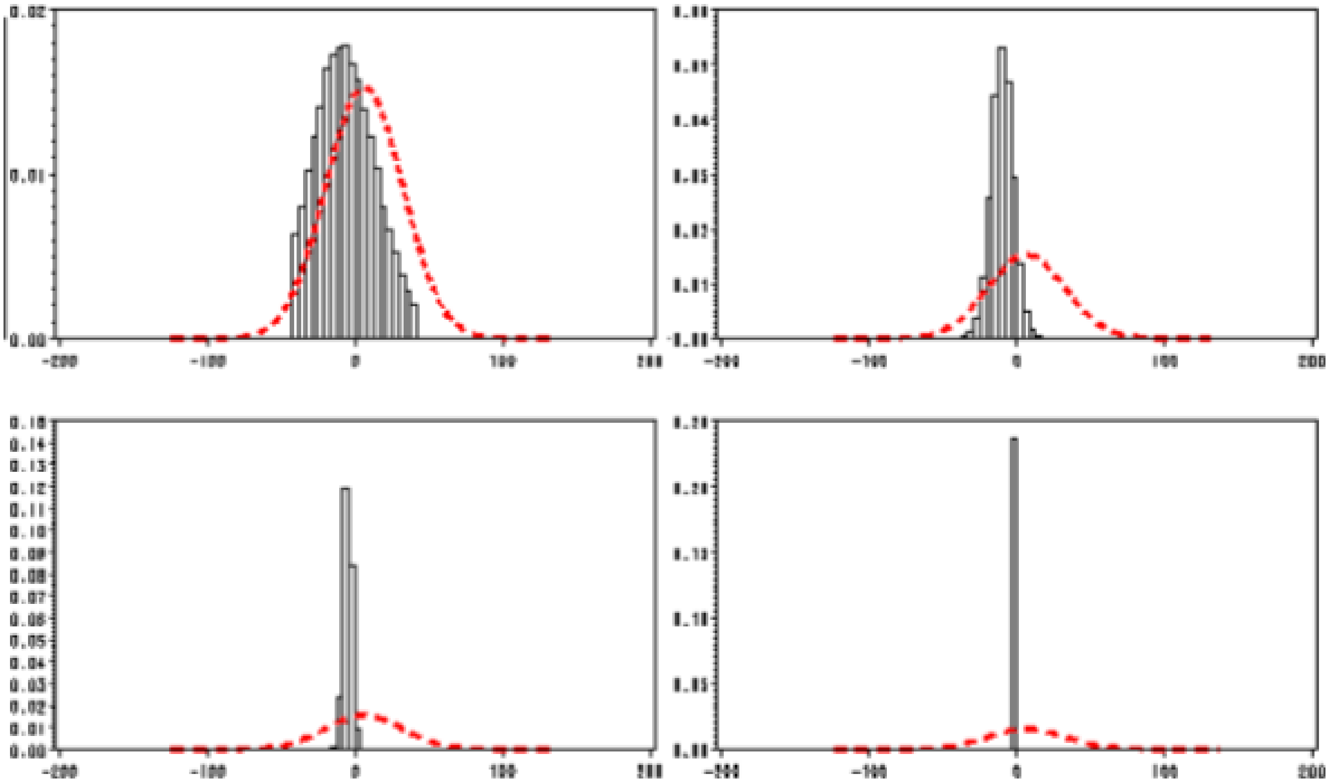

Figure 4.
Plots of the Prior (smooth line) and Posterior (histogram) distribution of the ITT estimates for each principal strata for the heterogeneous variance model.
Note: Plots from left to right and top to bottom are: Compliant mediating, Always mediating, Never mediating, and Defiant mediating.

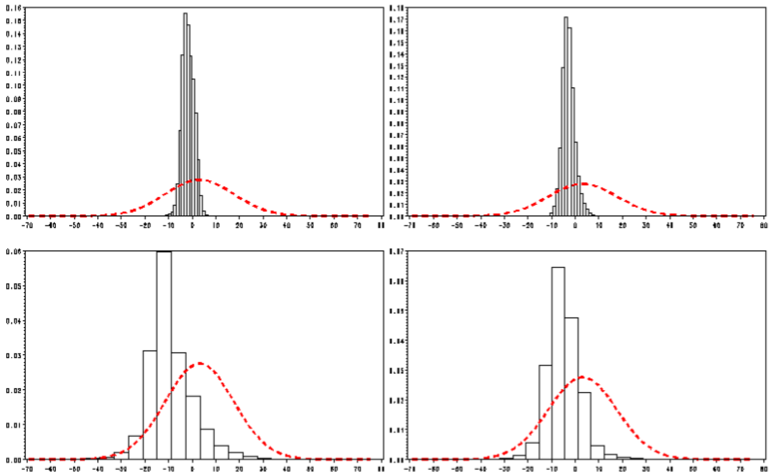

Figure 5.

Plots of the Prior (smooth line) and Posterior (histogram) distribution of the ITT estimates for each principal strata for the homogeneous variance model.

Note: Plots from left to right and top to bottom are: Compliant mediating, Always mediating, Never mediating, and Defiant mediating.

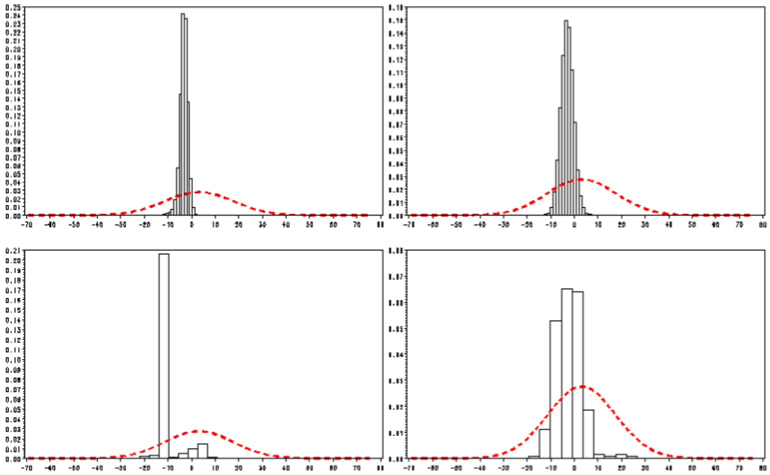

Figure 6.

Plots of the Prior (smooth line) and Posterior (histogram) distribution of the ITT estimates for each principal strata for the heterogeneous variance model.

Note: Plots from left to right and top to bottom are: Compliant mediating, Always mediating, Never mediating, and Defiant mediating.

Table 1

Performan ce of simulated models for CBT treatment for Suicide

Simulated Based on Heterogeneity for Defiers compared to other 3 groups					
	True Value	**Coverage**	**Mean Bias**	**Pct Bias**	**MSE**
Direct Effect					
Heterogeneous by Defiers Model	-5.70	96.2% 95.3%	-0.38 -0.06%	-6.6% -1.1%	8.07 3.91
Homogeneous Model	-5.70	97.4% 97.8%	-0.60 -0.30	-10.4% -5.2%	9.80 5.46
Direct Effect within Always Mediators					
Heterogeneous by Defiers Model	-10.05	97.6% 97.9%	1.89 0.76	18.7% 7.6%	47.22 24.32
Homogeneous Model	-10.05	97.6% 99.8%	5.06 4.38	50.1% 43.3%	31.30 29.90
Direct Effect within Never Mediators					
Heterogeneous by Defiers Model	-4.96	95.0% 95.3%	-0.79 -0.24	-15.8% -4.8%	9.65 4.66
Homogeneous Model	-4.96	95.0% 95.5%	-1.63 -1.24	-32.5% -24.9%	11.83 6.56

Mis-specifed Model - Simulation based on Heterogeneity for Always Mediators compared to other 3 groups					
	True Value	**Coverage**	**Mean Bias**	**Pct Bias**	**MSE**
Direct Effect					
Heterogeneous by Defiers Model	-5.70	96.4% 92.8%	0.74 0.74	12.9% 12.9%	10.09 5.46
Homogeneous Model	-5.70	95.6% 94.4%	0.69 1.24	12.0% 21.6%	13.16 9.50
Direct Effect within Always Mediators					
Heterogeneous by Defiers Model	-10.05	47.4% 36.7%	15.53 16.50	153.7% 163.3%	39.12 25.36
Homogeneous Model	-10.05	80.8% 78.9%	10.78 11.65	106.8% 115.3%	39.38 39.37
Direct Effect within Never Mediators					
Heterogeneous by Defiers Model	-4.96	92.8% 89.9%	-1.42 -1.57	-28.4% -31.4%	11.81 5.93
Homogeneous Model	-4.96	95.2% 93.2%	-0.67 -0.03	-13.5% -0.7%	15.13 10.46

Note: Per row top number corresponds to results for N=101 and bottom corresponds to results for N=202.

Table 2

Prior distribution assumptions for CBT treatment for Suicide

Parameters	Prior distribution
$\boldsymbol{\beta}_t$	N ($^{\square}\boldsymbol{\beta}$,Σ) with $\Sigma = nVar\,(^{\square}\boldsymbol{\beta})$
σ^2 under homogeneity of variance	Inverse-Gamma with parameters 0.01 and 0.01
σ_t^2 under heterogeneity variance for all t (t=1,2,3,or 4)	Each are Inverse -Gamma with parameters 0.01 and 0.01
δ_t	Dirichlet with parameters 1, 1, 1, 1

Table 3

Estimates of Intervention effect for CBT compared to TAU. Standard errors and nominal 95% confidence intervals are in parentheses

Effect		
Overall ITT effect	-6.35 (2.53) (-11.37,-1.33)	
Standard Direct Effectfor overall sample	-6.86 (2.60) (-12.01, -1.70)	
	Homogeneous Variance PS Model	**Heterogeneous Variance by Defiers PS Model**
Always mediators (12.9%)	-8.29 (17.68) (-43.43, 28.30)	-10.05 (7.63) (-24.92, 5.07)
Never mediators (75.2%)	-7.06 (4.47) (-15.36, 1.93)	-4.96 (2.80) (-10.43, 0.58)
Pooled Direct Effect of Intervention	-7.11 (4.28) (-15.16, 1.49)	-5.70 (2.65) (-10.92, -0.49)

Note: Principal strata probability per class under heterogeneous variance PS model are included in parentheses in the effect column

Table 4
Prior distribution assumptions for Depression specialist practice study

Parameters	**Prior distribution**
$\boldsymbol{\beta}_t$	$N(^{\square}\boldsymbol{\beta},\Sigma)$ with $\Sigma = nVar\ (^{\square}\boldsymbol{\beta})$
σ^2 under homogeneity of variance	Inverse-Gamma with parameters 0.01 and 0.01
σ_t^2 under heterogeneity variance for all t (t=1,2,3,or 4)	Each are Inverse-Gamma with parameters 0.01 and 0.01
δ_t	$N(0,diag(1,2.25,0.04)$)

Table 5

Estimates of Intervention effectfor practices with Depression specialist compared to those without Depression Specialist. Standard errors and nominal 95% confidence intervals are in parentheses

Effect		
Overall ITT effect	-3.12 (0.82) (-4.72,-1.51)	
Standard Direct Effect for overall sample	-2.67 (0.89) (-4.41, -0.93)	
	Homogeneous Variance PS Model	**Heterogeneous Variance by Defiers PS Model**
Always mediators (34.4%)	-2.83 (2.57) (-7.51, 3.10)	-2.88 (2.63) (-7.95, 2.34)
Never mediators (3.9%)	-9.17 (9.41) (-25.04, 13.73)	-8.52 (4.63) (-10.01, 4.89)
Pooled Direct Effect of Intervention	-3.83 (2.65) (-8.27, 2.15)	-3.38 (2.46) (-8.16, 1.46)

Note: Principal strataprobability for each strata under heterogeneous variance PS model are included in parentheses in the effect column