Published in final edited form as:

Stat Med. 2018 March 30; 37(7): 1149–1161. doi:10.1002/sim.7572.

A Bayesian Semiparametric Latent Variable Approach to Causal Mediation

Chanmin Kim^{a,*}, Michael J. Daniels^b, Yisheng Li^c, Kathrin Milbury^d, and Lorenzo Cohen^d

^aDepartment of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115

^bDepartment of Statistics, University of Florida, Gainesville, FL 32611

^cDepartment of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

^dDepartment of Palliative, Rehabilitation, and Integrative Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

Abstract

In assessing causal mediation effects in randomized studies, a challenge is that the direct and indirect effects can vary across subjects due to different measured and unmeasured characteristics. In that case, the population effect estimated from standard approaches implicitly averages over and does not estimate the heterogeneous direct and indirect effects. We propose a Bayesian semiparametric method to estimate heterogeneous direct and indirect effects via clusters, where the clusters are formed by both individual covariate profiles and individual effects due to unmeasured characteristics. These cluster-specific direct and indirect effects can be estimated through a set of regression models where specific coefficients are clustered by a stick-breaking prior. To let clustering be appropriately informed by individual direct and indirect effects, we specify a data-dependent prior. We conduct simulation studies to assess performance of the proposed method compared to other methods. We use this approach to estimate heterogeneous causal direct and indirect effects of an expressive writing intervention for patients with renal cell carcinoma.

Keywords

Bayesian nonparametrics; Causal inference; Cluster-specific effects; Effect modification; Sequential ignorability

*Correspondence : Chanmin Kim (chanmink@bu.edu), Department of Biostatistics, Boston University School of Public Health (current affiliation), Boston, MA 02118.

Software

Software in the form of R code is available on the GitHub repository (<https://github.com/lit777/LatentMediation>).

Supplementary Material

Supplementary material (including identification of cluster-specific effects, Gibbs sampling algorithm, postprocessing steps, additional simulation studies and descriptions of covariates in the data example) is available in the Web Appendix.

1 Introduction

Researchers often want to examine the extent to which an intervention has an impact on an outcome that is mediated by an intermediate variable. The effect of intervention that is through the intermediate variable, so-called *mediator*, has been defined as the indirect effect.^{1, 2} We define the direct effect as the effect on the outcome that is not attributable to changes in the mediator. Then, the total effect is the sum of the two effects.

Over the past decades, many approaches have been proposed for estimating the indirect, or mediation effect. Among many others, Baron and Kenny's regression models³ have been widely used in the behavioral and social sciences. Generally, this approach specifies three regression models for the outcome and mediator:

$$\begin{aligned} Y &= \alpha_1 + \beta_1 X + \varepsilon_1 \\ M &= \alpha_2 + \beta_2 X + \varepsilon_2 \\ Y &= \alpha_3 + \beta_3 X + \gamma M + \varepsilon_3, \end{aligned} \quad (1)$$

where either the first or the second equation is redundant.⁴ Then, the effects of interest can be estimated with products of coefficients; for example, if the second and third equations are considered, the total, indirect and direct effects are estimated by $\hat{\beta}_3 + \hat{\beta}_2 \hat{\gamma}$, $\hat{\beta}_2 \hat{\gamma}$ and $\hat{\beta}_3$, respectively, with $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\gamma}$ being the estimates of the corresponding parameters. MacKinnon et al.⁵ clearly describe several methods based on regression models.

The potential outcomes framework⁶ has also shed light on how one can draw causal inference from the observed data model in (1). First, to utilize the potential outcomes framework under 'Stable Unit Treatment Value Assumption (SUTVA)'⁷ in mediation analysis, one needs to specify an additional assignment mechanism for the intermediate variable given each treatment.^{8, 9} Given that, one can assume sequential ignorability⁴ which states that the mediator is ignorable after conditioning on the treatment and a set of confounders. Instead of conducting a formal mediation analysis, one can attempt to assess the mediation effect indirectly using principal stratification analysis.^{10–14} Mealli and Mattei,⁹ and VanderWeele^{15, 16} provide the details of these two types of analyses and investigate their relationships.

Motivated by the Baron and Kenny regression models, our Bayesian approach also consists of two regression models. Although Yuan and MacKinnon¹⁷ proposed a Bayesian structural equation model, our model differs in two ways: 1) we use a Bayesian profile regression method^{18, 19} which enables us to form clusters of covariate profiles flexibly; and 2) our target estimands are cluster-specific effects rather than the population effects. Additionally, our method can estimate the effects for each cluster which is formed not only by individual covariate profiles (or effect modifiers), but also by individual heterogeneous effects not explained by those observed covariate profiles, via a latent variable that depends on both individual effects and covariate profiles. Note that effect modification in our paper means effect modification *in measure* μ , i.e., $\mu(E(Y|X=1, Q=q), E(Y|X=0, Q=q))$ varies with strata Q .²⁰ Therefore, two subjects having the same covariate profiles of interest can be

assigned to two different clusters if their unobserved (hidden) traits drive the different effects. The use of profile regression allows us to model the effects of a number of potential effect modifiers (hypothesized by the investigators) by circumventing potentially difficult specification of the appropriate parametric form of the effects as well as dealing with the often present collinearity among them. In addition, the profile regression may also help simplify the interpretation of the heterogeneous covariate effects.

To let clustering be informed not only by the covariate profiles, but also by the additional hidden covariates, we specify an informative prior on the coefficient which is linked to the direct effect (β_3 in Equation 1). Forastiere et al.²¹ propose methods to estimate the cluster effects of subpopulations. However, their approach is more geared toward estimating *interference* between units within each ‘given’ cluster while our approach is focused on clustering itself and estimating the cluster-specific effects. **To evaluate heterogeneous effects (or effect modification) in mediation analysis, the analysis is typically conducted by including interaction terms between treatment and effect modifiers and/or between treatment and mediators in the models. This approach is sometimes referred to as “moderated mediation”.²² Preacher et al.²³ provides a thorough guide to estimating for moderated mediation and to test a variety of hypotheses. Muthén and Asparouhov²⁴ also show that moderated mediation effects can be expressed as functions of effect modifiers and can be evaluated at different values. However, this approach of adding interaction terms cannot provide information on latent clusters of covariate profiles (or effect modifiers) on which those effects are defined.** Thus, to the best of our knowledge, there has not been any approach developed for estimating heterogeneous effects for the mediation based on clusters of covariate profiles (or effect modifiers).

We apply our method to a randomized controlled trial of expressive writing for patients with renal cell carcinoma (RCC).²⁵ The goal of this trial was to evaluate the effect of interventions (expressive writing, EW) on quality-of-life. Patients in the study completed the Center for Epidemiologic Studies Depression (CES-D), MD Anderson Symptom Inventory (MDASI), Brief Fatigue Inventory (BFI), Pittsburgh Sleep Quality Index (PSQI), and Impact of Event Scale (IES) at baseline and 1, 4, and 10 months after the intervention. Here, we examine the IES, a 15-item scale assessing thought intrusion and avoidance during the prior week, measured at 1 month as a potential mediator and the 10 month MDASI, which asks patients to rate the severity of 13 core symptoms common across all cancer diagnoses and treatments and the extent to which these symptoms interfere with daily activities, as the outcome. Higher scores correspond to greater severity and interference. Several baseline covariates were also available, some of which are hypothesized as potential effect modifiers with others being potential confounders, thus considered the basis for the covariate profile of interest.

The remainder of this article is organized as follows. In Section 2, we propose the models for the mediator and outcome and the model for assigning individuals to clusters. Section 3 defines the cluster-specific direct, indirect and total effects and introduces an informative prior for the direct effects. Section 4 outlines posterior inference. In Section 5, we conduct simulation studies to assess performance of the proposed model. In Section 6, we apply our

method to the RCC data and examine cluster-specific mediation effects. Finally, Section 7 summarizes the paper and proposes potential extensions.

2 Framework on Latent Mediation Model

2.1 Notation and Models

We denote X_i , M_i and Y_i as the observed treatment, mediator, and outcome of individual $i = 1, \dots, n$, respectively. Then, we define a covariate profile for individual i as $\mathbf{u}_i = (u_{i1}, u_{i2}, \dots, u_{ip})$ where u_{ij} are the individual covariates. Covariate profiles are clustered into groups identified by an allocation variable, $Z_i = k$, $k = 1, \dots, K$, indicating the cluster to which the i -th individual belongs. The mediator and outcome models for each $i \in \{1, \dots, n\}$ are defined as follows,

$$M_i = \alpha_{0,Z_i} + \alpha_{1,Z_i} X_i + \varepsilon_{Z_i,i}, \quad (2)$$

$$Y_i = \alpha_{2,Z_i} + \alpha_{3,Z_i} X_i + s(M_i, \boldsymbol{\beta}_{Z_i}) + \zeta_{Z_i,i}, \quad (3)$$

where coefficients α_{0,Z_i} , α_{1,Z_i} , α_{2,Z_i} , α_{3,Z_i} and $\boldsymbol{\beta}_{Z_i}$ vary by clusters ($Z_i = k$) and the error terms are specified independently as $\varepsilon_{k,i} \sim N(0, \sigma_k^2)$ and $\zeta_{k,i} \sim N(0, \tau_k^2)$.

The mediator, M_i , is entered into model (3) via a smooth function, $s(\cdot, \boldsymbol{\beta}_{Z_i})$ to decrease the change of mis-specification. We use low-rank thin-plate penalized splines for computational efficiency.²⁶ Formally, the low-rank thin-plate spline representation of $s(\cdot, \boldsymbol{\beta}_{Z_i})$ for a one-dimensional variable^{26, 27} is

$$s(M_i, \boldsymbol{\beta}_{Z_i}) = \beta_{0,Z_i} M_i + \sum_{j=1}^J \beta_{j,Z_i} |M_i - \xi_{j,Z_i}|^3,$$

where knots ξ_{j,Z_i} , $j = 1, \dots, J$ can be determined by the sample quantiles of the mediator in cluster Z_i and $\boldsymbol{\beta}_{Z_i}^\top = \{\beta_{0,Z_i}, \beta_{1,Z_i}, \dots, \beta_{J,Z_i}\}$. After some reparametrization, the outcome model in (3) can be re-expressed as

$$Y_i = \alpha_{2,Z_i} + \alpha_{3,Z_i} X_i + \beta_{0,Z_i} M_i + \mathbf{T}_{Z_i}(M_i) \mathbf{f}_{Z_i} + \zeta_{Z_i,i}, \quad (4)$$

where $\mathbf{f}_k = \Omega^{1/2} \mathbf{k} \beta_k^0$, $\beta_k^0 \equiv \beta_k \setminus \beta_{0,k} = \{\beta_{1,k}, \dots, \beta_{J,k}\}^\top$ and the thin-plate spline penalty matrix Ω_k has (j, h) -th entry $|\xi_{j,k} - \xi_{h,k}|^3$. Also,

$T_k(M_i) = \{|M_i - \xi_{1,k}|^3, \dots, |M_i - \xi_{J,k}|^3\} \Omega_k^{-1/2}$. Finally, we set $\mathbf{f}_k \sim MVN(\mathbf{0}, \sigma_k^f \mathbf{I}_J)$ and $\sigma_k^f = \lambda_k \tau_k^2$. See the Web Appendix A for more details.

Since there may be some empty clusters, for each cluster k , we set

$\alpha_{0,k} | \sigma_k^2 \sim N(0, \sigma_k^2 \times \text{Var}(M))$, $\alpha_{1,k} | \sigma_k^2 \sim N(0, \sigma_k^2 \times \text{Var}(M))$ and $\alpha_{2,k} | \tau_k^2 \sim N(0, \tau_k^2 \times \text{Var}(Y))$, $\beta_{0,k} | \sigma_k^2 \sim N(0, \sigma_k^2 \times \text{Var}(Y))$, where the variance of the observed data, $\text{Var}(\cdot)$, provides some information for the empty clusters. We specify a Gamma(a, b) prior on $1/\sigma_k^2$ and $1/\tau_k^2$ with mean parameter a/b . From empirical trials we found that $a = N/20$ and $b = N/40$ (N = total sample size) are reasonable default settings. In Section 3.2, we introduce a data-dependent prior for the coefficient α_{3, z_i} .

In the next subsection, we introduce the model that allocates each individual $i \in 1, \dots, n$ to a cluster $k \in \{1, \dots, K\}$.

2.2 Cluster Assignment Model

Following Molitor et al.,¹⁸ we assume the P covariates, $\mathbf{u} = (u_1, u_2, \dots, u_P)$, are categorical (below we describe the extension to continuous covariates) with n_p categories for the p th covariate. These are the *effect modifiers* in our framework. Denote ψ_k as the probability of assignment to the k -th cluster, and let $\phi_k^p(u)$ denote the probability that the p -th covariate in cluster k is equal to u . In other words, for cluster k , the parameters, ϕ_k^p , $p = 1, \dots, P$, define the prototypical profile for that cluster. Then, the basic mixture model for assignment is

$$\begin{aligned} Pr(\mathbf{u}) &= \sum_{k=1}^K Pr(Z=k) \prod_{p=1}^P Pr(u_p | Z=k) \\ &= \sum_{k=1}^K \psi_k \prod_{p=1}^P \phi_k^p(u_p), \end{aligned} \quad (5)$$

where the mixture weights correspond to a maximum of K clusters, with K being set to a sufficiently large number to capture all possible clusters. The weights are denoted as $\Psi = \{\psi_k, k = 1, \dots, K\}$, and are specified in terms of a “stick-breaking” prior.^{28, 29} In particular, $V_1, \dots, V_{K-1}$ are distributed as independent Beta(1, γ) and $\psi_k = V_k \prod_{h < k} (1 - V_h)$ for $k = 1, \dots, K-1$ with $V_K = 1$ to approximate the stick-breaking construction with K clusters. We place a Gamma(1, 1) prior on the parameter γ and place a symmetric Dirichlet prior on ϕ_k^p for each (k, p) pair with concentration parameter $\boldsymbol{\lambda}_{(p)} = \{\lambda_{1,(p)}, \dots, \lambda_{n_p,(p)}\}$.

As mentioned above, this approach can be extended to non-categorical covariates. With a continuous covariate, u_p , we can specify $p(u_p | Z=k) = \phi(u_p; \mu_k^p, \sigma_k^p)$ where $\phi(\cdot; a, b)$ is a normal density with mean a and variance b , $\mu_k^p \sim N(\mu_0, \kappa_0 \sigma_k^p)$, and $1/\sigma_k^p \sim G(\alpha_\sigma, \beta_\sigma)$. In Section 5 we demonstrate model performance in the setting of both categorical and continuous covariates and discuss the choice of the hyper-parameters.

We can also incorporate additional covariates into models (2) and (4) to adjust for confounding between the outcome and the mediator. Also, there can be differential confounding relationships within each cluster. We revisit these issues in Section 3.1

3 Cluster-Specific Effects

Based on the specification in the previous section, we define cluster-specific direct and indirect effects. The direct effect of cluster k , DE_k , is defined as the coefficient of the treatment X in model (3). That is, the cluster-specific direct effect for individuals who belong to cluster k is $DE_k = \alpha_{3, k}$. The indirect effect of cluster k , IE_k , which is the effect that is attributable to the mediator, is defined as $IE_k = \alpha_{1, k}\beta_{0, k} + T_k(E[M(1)])f_k - T_k(E[M(0)])f_k$ where $T_k(E[M(1)])$ and $T_k(E[M(0)])$ denote $T_k(M)$ from model (4) with M_j being replaced by $E(M_j)$ evaluated at $X_j = 1$ or 0 based on model (2), respectively. Since the total effect of cluster k , TE_k , is the effect of the treatment on the outcome accounting for all causal paths (passing and bypassing the mediator) for individuals assigned to cluster k , it is defined as $TE_k = IE_k + DE_k$.

To interpret these cluster-specific effects as valid causal effects, we utilize the potential outcome framework⁶ described in the next section.

3.1 Potential Outcome Framework and Natural Effects

Before reviewing the potential outcome framework, we first provide details on SUTVA⁷ which states that there is ‘no interference’ between units and that there is ‘no hidden version of treatments’. In mediation analysis, ‘no hidden version of treatments’ can be conceptualized by formulating the assignment mechanism for the mediator within each treatment level.³⁰ The most widely used one is sequential ignorability⁴ that assumes a randomized mediator conditioning on the treatment and confounders. Now, let $M_i(x)$ denote the potential value of the mediator that would have been observed had the treatment for the i -th individual, X_i , been set to level x . Similarly, $Y_i(x, m)$ represents the potential value of the outcome that would have been observed had treatment X_i been set to level x , the mediator for the i -th individual, M_i , set to value m . Then, the observed outcomes for individual i are $M_i^{\text{obs}} = M_i(X_i)$ and $Y_i^{\text{obs}} = Y_i(X_i, M_i(X_i))$. For individual i , under the opposite treatment $1 - X_i$, the potential mediator $M_i(1 - X_i)$ and outcome $Y_i(1 - X_i, M_i(1 - X_i))$ are never observed. In addition, there is an *a priori* counterfactual outcome $Y_i(x, m')$ for m'

$M_i(x)$ that is never observable. There are sometimes concerns about plausibility of the sequential ignorability and many approaches have been proposed to assess sensitivity to this assumption including Imai et al.,⁴ VanderWeele,³¹ Tchetgen Tchetgen and Shpitser,³² and Kim et al.³³

Within the potential outcomes framework described above, we assume sequential ignorability, similar to Imai et al.,⁴ but for each cluster k

$$\{Y_i(x', m), M_i(x)\} \perp X_i | Z_i = k, W_i = w \quad (6)$$

$$Y_i(x', m) \perp M_i | X_i = x, Z_i = k, \mathbf{W}_i = \mathbf{w}, \quad (7)$$

for $x, x' = 0, 1$, where $\mathbf{W}_i$ is a vector of baseline covariates to adjust for confounding relationships between the outcome and mediator. If the data are not randomized within each cluster k , $\mathbf{W}_i$ adjusts for confounding relationships among the treatment, mediator and outcome within that cluster. As with the other coefficients in models (2) and (4), we allow the coefficients of $\mathbf{W}_i$ in those models to vary by clusters. Thus, the full set of confounders is the same across clusters but their confounding relationships may differ through the coefficients. The assumption that $Pr(X_i = x | Z_i = k, \mathbf{W}_i = \mathbf{w}) > 0$ and $p(M_i = m | X_i = x, Z_i = k, \mathbf{W}_i = \mathbf{w}) > 0$ for all $x, m, k, \mathbf{w}$ values (i.e., positive pmf and pdf for the treatment and mediator for all values, respectively) is also required. Finally, we assume no unmeasured post-treatment confounder. Note that we assume all confounders in (6) and (7) are also included in the set of the potential effect modifiers (i.e., $\mathbf{W} \subset \mathbf{u}$).

Now, we can define the causal effects within clusters. To simplify notation, we suppress the index i . The natural direct effect of cluster k is defined as $NDE_k = E[Y(1, M(0)) - Y(0, M(0)) | Z = k]$. This quantifies the effect of the treatment $X = 1$ on the outcome where the mediator is set to its 'natural' value $M(0)$, which would have been observed had treatment X been set to 0. The natural indirect effect of cluster k is defined as $NIE_k = E[Y(1, M(1)) - Y(1, M(0)) | Z = k]$, i.e., the treatment effect through the change in the mediator had everyone received treatment $X = 1$. Under the aforementioned sequential ignorability (6)–(7) and models (2) and (4), these effects can be estimated and are functions of the coefficient estimates for cluster k (see the Web Appendix). In particular, the estimate of the natural direct effect of cluster k is given by $\widehat{NDE}_k = \hat{\alpha}_{3,k}$ and the natural indirect effect of cluster k is estimated by $\widehat{NIE}_k = \hat{\alpha}_{1,k} \hat{\beta}_{0,k} + \hat{T}_k(E[\hat{M}(1)]) \hat{f}_k - \hat{T}_k(E[\hat{M}(0)]) \hat{f}_k$. The proof is given in the Web Appendix B.

3.2 A Data-dependent Prior for the Direct Effects

In our framework, the outcome and assignment models are fitted jointly. Therefore, information from the outcome models informs the latent allocating variable Z which in turn influences the parameter ψ_k in (5). Thus, clustering is also informed by variation in the indirect and direct effects. That is, our proposed model determines the clusters based on individual effect through values of the outcome and the mediator as well as individual covariate profiles. As a result, two individuals who have the same covariate profile still can be assigned to two different clusters. This is possible when there are some unmeasured covariates that lead to different effects through values of the outcome and the mediator. In this way, we can leverage our method to find cluster-specific effects induced from sources beyond the measured characteristics.

To let clustering be sufficiently informed by (2) and (4), we specify an informative prior on the coefficient $\alpha_{3,k}$ in the outcome model (4). Without loss of generality, we assume the maximum number of clusters is an odd number K and denoting $K' = K - 1$, we set normal priors on $\alpha_{3,k}$ for all $k = 1, \dots, K$ as follows,

$$\alpha_{3,k} \sim N(\chi_k, s^2). \quad (8)$$

χ_k has a multinomial distribution with K possible values, Λ , equally spaced from $\chi - \delta$ to $\chi + \delta$,

$$\Lambda = \left\{ \chi - \frac{K'/2 \times \delta}{K'/2}, \chi - \frac{(K'/2-1) \times \delta}{K'/2}, \dots, \chi - \frac{1 \times \delta}{K'/2}, \chi, \chi + \frac{1 \times \delta}{K'/2}, \dots, \chi + \frac{K'/2 \times \delta}{K'/2} \right\} \quad (9)$$

with corresponding probabilities $\mathbf{P}_\Lambda = \{p_{\Lambda,1}, \dots, p_{\Lambda,K}\}$. $\mathbf{P}_\Lambda$ is given a Dirichlet prior. The center value of $\mathbf{P}_\Lambda$, χ , is set to the estimate of $\alpha_{3,k}$ from fitting model (4) to the entire data without clustering. The sensitivity parameter δ determines the range of possible mean values on both sides of the population effect χ . We assume the deviation of any cluster-specific effect from the population effect does not exceed $2 \times SD(Y[X=1] - Y[X=0])$ (i.e., two standard deviations of the total effect) on both sides. This is based on an assumption that any cluster-specific ‘direct’ effect is within two standard deviations of the population ‘direct’ effect, where two standard deviations of the population ‘total’ effect encompasses this range. However, if more variability of any cluster-specific effect is expected, we can increase the range. Since for any individual we never observe potential outcomes under both treatments simultaneously, we approximate $SD(Y[X=1] - Y[X=0])$ with a conservative bound

(assuming a positive correlation), $\sqrt{Var(Y[X=1]) + Var(Y[X=0])}$ where $Var(Y[X=1])$ and $Var(Y[X=0])$ are the empirical variances of the observed outcomes. We equate the standard deviation parameter in (8), s to the distance between the adjacent mean parameters in (9), $\delta(K'/2)$ to control the magnitude of overlap between priors for distinct mean parameters, $\chi_k - \chi_h$ and the minimum difference in α_{3k} ’s between clusters. For example, if $s = \frac{1}{K'}\delta$, then any two adjacent mean parameters are separated by 2 standard deviations. Note that the results are not very sensitive to choices of K since the standard deviation s in (8) changes concurrently with K .

The overall rationale behind this prior specification is that we allow each $\alpha_{3,k}$ to have its own prior with small variance. The sensitivity parameters δ and s control the extent to which each prior differs. Since the sensitivity parameter δ determines both the range of Λ and the difference between adjacent mean parameters in (9) ($\frac{2}{K'}\delta$), this parameter should be determined by the clinically meaningful difference in a study. For the standard deviation parameter s , we can consider $n_s \times s = \frac{2}{K'}\delta$ for $n_s = 1, 1.5, 2$, i.e., the mean parameters are separated by at least 1, 1.5 and 2 standard deviations, respectively. We examine sensitivity to this specification in the simulation study (Section 5).

4 Posterior Inference

The cluster-specific effects are functions of $\Gamma = \{\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \boldsymbol{\beta}_0, f, \boldsymbol{\psi}, \boldsymbol{\phi}\}$ where a bold character denotes a vector of parameters for all K clusters. Bayesian inference for the cluster-specific effects is based on the following likelihood:

$$L(\mathbf{Y}, \mathbf{M}, \mathbf{X}, \mathbf{W}, \mathbf{u} | \Gamma) \propto \sum_{\mathbf{Z}} \prod_{i=1}^n \left\{ Pr(Y_i | X_i, M_i, W_i, Z_i) Pr(M_i | X_i, W_i, Z_i) Pr(Z_i) \prod_{p=1}^P Pr(u_{p,i} | Z_i) \right\},$$

where the components in the brackets are specified in models (2), (4) and (5). This representation explicitly shows that clustering is informed by all components. Coupled with a prior distribution $p(\Gamma)$, the draws from the resulting posterior are plugged into functions that define the cluster-specific effects to obtain posterior distributions of these.

An R package `PREMIUM` is available for Bayesian profile regression.³⁴ However, it cannot handle two models (for the outcome and the mediator) simultaneously and only allows the intercept term to vary by clusters with no other coefficients varying by clusters. For these reasons, we wrote our own R code for the MCMC algorithm. See Web Appendix C for details.

4.1 Postprocessing Steps

In mixture models, posterior samples may not represent each cluster well due to the “label switching” problem. To overcome this issue, many methods have been proposed such as the relabeling algorithm.³⁵ In Dirichlet process mixture models (DPM), this problem is less of an issue since labels are constrained by the stick-breaking construction of the mixing parameters. If label switching occurs, however, it is difficult to handle the issue by using general methods proposed for the mixture model since, in every iteration, switching between dimensions of clusters could occur in addition to the label switching. To overcome this, we follow the scheme from Molitor et al.¹⁸ and Liverani et al.³⁴ construct the posterior similarity matrix S and then find the best partition using partitioning around medoids (PAM)³⁶ where medoids are similar to centroids or means, but they are always elements of the data set. See Web Appendix D for details.

Once the best partition $\mathbf{P} = \{P_1, \dots, P_K\}$ is obtained, we identify the subjects that belong to each P_k indexed by $i_1, P_k, i_2, P_k, i_3, P_k, \dots, i_{n_k}, P_k$. Then, we repeat the following steps after an appropriate burn-in:

1. From the m -th MCMC sample, extract the sampled allocation variables Z_i for subjects belong to P_k (i.e., $i_1, P_k, i_2, P_k, i_3, P_k, \dots, i_{n_k}, P_k$) and denote them as $Z_{i_1, P_k}^{(m)}, Z_{i_2, P_k}^{(m)}, \dots, Z_{i_{n_k}, P_k}^{(m)}$.
2. Find unique values of allocation variables $Z_{i_1, P_k}^{(m)}, Z_{i_2, P_k}^{(m)}, \dots, Z_{i_{n_k}, P_k}^{(m)}$ and denote them as $Z_1, P_k, Z_2, P_k, \dots, Z_u, P_k$ where $u = i_{n_k}$.

3. From the m -th MCMC sample, extract the parameters $\alpha_{0,h}, \alpha_{1,h}, \alpha_{2,h}, \alpha_{3,h}, \beta_{0,h}, f_h$ corresponding to all $h \in \{Z_{1,P_k}, Z_{2,P_k}, \dots, Z_{u,P_k}\}$.
4. Average these samples over $h, \{\bar{\alpha}_{0,h}, \bar{\alpha}_{1,h}, \bar{\alpha}_{2,h}, \bar{\alpha}_{3,h}, \bar{\beta}_{0,h}, \bar{f}_h\}$, and define these as the new m -th parameter samples for all subjects in P_k .
5. Repeat 1–3 for every $P_k \in P$ for each m .

5 Simulation Studies

We assess the performance of the proposed model based on a simulation study under sequential ignorability where we incorporate two covariates into the outcome and mediator models to represent confounding relationships between treatment and mediator and between treatment and outcome. We consider four different models: (I) a standard frequentist model without clustering; (II) a standard frequentist model + clustering covariates by the k -means algorithm; (III) a standard frequentist model + clustering covariates, outcomes and mediators by the k -means algorithm; (IV): our Bayesian profile regression model. For Models (I–III), we use Generalized Additive Models (GAM)³⁷ (with a cubic spline for the mediator term in the outcome model) for the mediator and the outcome and use the R package `mediation` (CRAN)³⁸ to fit them. In Models (II) and (III), the k -means clustering algorithm is used to find the best clustering of the covariates (and the outcome and the mediator data) and then the outcome and mediator models (GAMs) are fitted in each cluster. For the proposed Bayesian profile regression model, we set $\alpha_\gamma = \beta_\gamma = 1$ and $\alpha_p = \{1, \dots, 1\}$ with the other hyper-parameters following the specification in Web Appendix C.

We examine five different scenarios: **Scenario 1** examines performances of the models with all confounders; **Scenario 2** examines performances of the models with one missing confounder (but that confounder is included in the pool of effect modifiers). In Web Appendix E, we examine the scenarios 3, 4 and 5: **Scenario 3** examines performances of the models with one missing confounder which is not included in the pool of effect modifiers; **Scenario 4** examines performances of the models with correlated effect modifiers; and finally **Scenario 5** examines performances of our model when there are no clusters.

In the simulated data ($n = 300$, similar to the original sample size of the data application in Section 6), we generate two continuous covariates and four categorical covariates for the effect modifiers, $u_1, u_2, u_3, u_4, u_5, u_6$, from distributions varying by the clusters (see Figure 1 and Table 1). The first two covariates, u_1 and u_2 , are also used as confounders. The binary treatment is generated from the conditional distribution $p(X = 1 | u_1, u_2) = \text{expit}(0.6 + 0.6u_1 - 0.5u_2)$. The outcome and the mediator are simulated to form three clusters whose sizes of which are different from those of clusters formed by the effect modifiers. The first 50 data points are generated from a normal outcome model with a nonlinear term of the mediator, the second 100 data points are generated from the normal outcome model with a linear term of the mediator, and the final 150 data points are generated from a skewed normal outcome model with a linear term of the mediator **to assess sensitivity to the normality assumption for the cluster-specific error terms in our models**. Therefore, in principle, there exist four distinct clusters with different sizes ($n_1 = 50, n_2 = 100, n_3 = 100, n_4 = 50$) (Figure 1 summarizes how the data are generated from two different levels and form the final four

clusters.): Cluster 1, Cluster 2, Cluster 3 and Cluster 4, where subjects assigned to Clusters 2 and 3 should have common covariate profiles and subjects assigned to Clusters 3 and 4 should have the same cluster-specific effects.

For the proposed model, the possible range of the mean parameters, Λ , of the informative prior for $\alpha_{3,c}$ is determined by $\delta = \sqrt{\text{Var}(Y|X=1) + \text{Var}(Y|X=0)} \approx 11$. Note that the k -means algorithm requires a fixed number of clusters while our proposed model only requires we specify an upper bound (K) for the number of possible clusters. We set $K = 21$ which is almost five times more than the truth. In general, a sufficiently large number of K with the mixture weight ψ_K of the cluster assignment model less than 0.01 can be chosen. However, since K determines the distance between the adjacent mean parameters in (9), it is recommended not to set K to an arbitrary large number which induces no clinically meaningful difference between the adjacent mean parameters.

To match the effect estimates to the true cluster-specific effects, we find the resulting partitions $\mathbf{P} = \{P_1, P_2, P_3, P_4\}$ of the data ($n = 300$) and consider that P_k represents ‘Cluster 1’ if observations from the data points $n = 1, \dots, 100$ appear in partition P_k the most often and consider that P_k represents ‘Cluster 2’ if observations from the data points $n = 101, \dots, 175$ appear in partition P_k the most often. Similarly, we can identify partitions for the remaining clusters. And the causal effects from each P_k are linked to the corresponding cluster. We do this step for both models (II) and (III). For our proposed model, we find the best partition as described in Section 5 and follow the exact same procedure except that the resulting partition can consist of more (or less) than 4 clusters, i.e., $\mathbf{P} = \{P_1, \dots, P_c\}$ where c can be any positive integer less than or equal to K . Also, note that there is the possibility that the same partition P_k represents more than one true cluster if P_k contains many observations from multiple clusters.

Table 2 summarizes the results based on 500 replications. It shows that our proposed model performs well in terms of biases and MSEs under scenarios 1 and 2. In Table 2, the estimates in bold indicate biases larger than 0.25 and MSEs larger than 0.5; these only appear in Models (I–III). The results for Model (I) show how much the cluster-specific effects are deviated from the population effects. Since Model (II) only uses the covariates (and not any outcome and mediator information) to form the clusters, the poor estimates were expected. However, even when we account for the outcome and mediator information in the clustering procedure in Model (III), it does not estimate the cluster-specific effects correctly and this is also expected as it is not ‘clustering’ the right quantities. Besides the clustering method, large biases in Model (II) and (III) are partly due to GAM since it does not perform well in the presence of non-standard distributions.³³ In Scenario 2, with the one missing confounder, our model still performs well since this missing confounder is still adjusted for, to some extent, by being a variable in the pool of potential effect modifiers. **When the data are generated from a non-normal distribution (Cluster 4), our model still performs well in terms of biases and MSEs.**

We repeated the same simulation study with a larger sample size, $n = 1500$. Table 2 shows that our model continues to perform well. Under both scenarios, the biases and MSEs from

our model are generally much smaller than those with the smaller sample size while the estimates from Models (II) and (III) do not improve with more data.

To check sensitivity to K , we also fit our proposed Bayesian profile regression model with $K = 15$ for Scenario 1. The MSEs slightly increase for clusters 2 and 3, but the biases almost stay the same (see Table 3).

In Web Appendix E, we include the results of scenarios 3, 4, and 5. In Scenario 3, two continuous covariates, U_1 and U_2 , are slightly correlated (0.04–0.05 for each cluster). In Scenario 4, we assume U_1 is unmeasured (i.e., missing confounder and effect modifier). Table S1 summarizes the results for scenarios 3 & 4. Under both, our model still shows stable performance while Models (II) and (III) estimates the cluster specific effects imprecisely. In Scenario 5, the data ($N = 300$) are generated based on the same scheme used to form Cluster 1 in Figure 1 (i.e., no clusters). Table S2 shows that our Bayesian profile regression model is able to capture a single cluster and estimate the homogeneous effects.

Finally, we assessed sensitivity to the standard deviation parameter s in (8). Table 2 shows the case of $n_s = 2$ (i.e., two adjacent mean parameters are separated by 2 standard deviations); however, results are similar for $n_s = 1.5$ (i.e., two adjacent mean parameters are separated by 1.5 standard deviations) (see Table 3) under Scenario 1.

6 Application

We apply our method to the randomized clinical trial enrolling patients suffering from renal cell carcinoma (RCC) to examine the effectiveness of expressive writing (EW) on quality of life.²⁵ Participants ($N = 277$) with stage I to IV RCC were randomized to one of two treatment arms: (1) expressive writing group in which subjects were supposed to write about their deepest thoughts and feelings ($N_1 = 138$); or (2) neutral writing group (NW) in which subjects were supposed to write about neutral topics on four separate occasions ($N_0 = 139$). Subjects' physical and emotional logs were recorded by MD Anderson Symptom Inventory (MDASI), Brief Fatigue Inventory (BFI), Center for Epidemiologic Studies Depression Scale (CES-D), Pittsburgh Sleep Quality Index (PSQI), Medical Outcomes Study Short Form-36 (SF-36), and Impact of Event Scale (IES) measured at baseline and 1, 4, and 10 months after randomization.

The 10-month MD Anderson Symptom Inventory (MDASI), which asks patients to rate the severity of 13 core symptoms common across all cancer diagnoses and treatments and the extent to which these symptoms interfere with daily activities was defined as the primary outcome. The Impact of Event Scale (IES), a 15-item scale assessing thought intrusion and avoidance during the prior week, measured at 1 month was examined as a potential mediator with higher scores representing greater intrusive thoughts and avoidance behaviors.

A variety of potential effect modifiers ($\mathbf{u}$) were collected including “BMI”, “Diabetes”, “Heart Condition”, “High Blood Pressure”, “Cancer other than Kidney Cancer”, “Smoking”, “CES-D”, “Treatment”, “Age”, “Gender”, “Stage”, “PreTreatment”, “Cell Type”, “STATE_{*b*}”, “PSQI_{*b*}”, “Surgery”, “MDASI_{*b*}” and “IES_{*b*}” where a subscript b indicates a baseline measurement. Details on these effect modifiers are in Table S3 in Web Appendix E.

Also, for the 36 subjects who have incomplete data for covariate profiles, we assume the missingness is ignorable.

6.1 Priors and Parameters

We specify $\sigma_k^2 \sim IG(8, 4)$ and $\tau_k^2 \sim IG(8, 4)$ for models (2) and (4) based on the specification in Section 2.1 and the total sample size. We restrict the possible range for the mean parameters in Λ , $2\delta = 2 \times \sqrt{\text{Var}(Y|X=1) + \text{Var}(Y|X=0)} = 3.9$. We consider a clinically meaningful difference in effect to be as small as 0.5 based on the consultation with experts from the original study; that is, we want the distance between the adjacent mean parameters in (9), $\delta(K'/2)$ to be larger than 0.5. Thus, the maximum number of clusters, $K (= K' - 1)$, should not exceed 8; here, we approximate the infinite DPM for the assignment model using 7 clusters. The number of knots ($\xi_{j,k}, j = 1, \dots, J; k = 1, \dots, 7$) in Model (4) is set to 5 (1/6th, 2/6th, 3/6th, 4/6th, 5/6th percentiles of mediators in each cluster k), since this produced the maximum adjusted R-squared value of the outcome model for the population among $J = 1, \dots, 10$ equally spaced knots.

6.2 Results

We run 30000 MCMC iterations and use the last 10000 iterations with thinning of 10 to obtain $M = 1000$ posterior samples. It is known that it can be difficult to check MCMC convergence of the DPM model,³⁹ and so we instead check the posterior of $Pr(Z)$, the marginal posterior distribution of the allocation probabilities. Based on visual inspection of the trace plots of four chains (not shown), convergence and mixing seemed good.

We present the indirect, direct and total effects for clusters that explain most of the variation in the covariate profiles and in the outcome and mediator. The best partition consists of three clusters (say $k = 1, 2, 3$) with corresponding assignment probabilities $Pr(Z = 1) = \psi_1 = 0.52$, $Pr(Z = 2) = \psi_2 = 0.25$ and $Pr(Z = 3) = \psi_3 = 0.23$. Table 4 contains the cluster-specific indirect, direct and total effects with the corresponding posterior means of the assignment probabilities. Figure 2 depicts differences in selected covariates (i.e., stage, treatment, and surgery) which show the most pronounced differences among the three clusters. Cluster 1 consists mainly of early stage patients who had been receiving a non-systemic treatment, while Cluster 3 consists mainly of late stage patients who had been receiving a systemic treatment.

In Cluster 1, the mean (and 95% C.I.) of the total effect on MDASI at 10 months is estimated as $-0.19(-0.77, 0.43)$, the indirect effect through the Impact of Event Scale (IES) is estimated as $-0.02(-0.54, 0.52)$ and the direct effect is estimated as $-0.18(-0.49, 0.13)$ while, in Cluster 2, the total effect on MDASI at 10 months is $-0.33(-1.60, 1.34)$, the indirect effect is estimated to be $0.12(-0.92, 1.66)$ and the direct effect is $-0.44(-1.17, 0.42)$. Finally, in Cluster 3, the total effect on MDASI at 10 month is $-0.15(-1.23, 0.69)$, the indirect effect is estimated to be $-0.28(-1.27, 0.34)$ and the direct effect is $0.13(-0.49, 0.71)$. Based on the results, the expressive writing appears to have more impact (between Cluster 1 and Cluster 3) on reduction in cancer-related symptoms (MDASI) for patients who had the

nonsystemic treatment (on average) and most of its effect comes from the pathway that is not through the IES.

To compare the cluster-specific effects to the population effects, we run the same outcome models without clustering ($K = 1$). The population total effect is estimated as $-0.37(-0.88, 0.15)$, and the population indirect and direct effects are estimated as $-0.18(-0.53, 0.12)$ and $-0.19(-0.60, 0.22)$, respectively. On average, the total effect estimate at the population level is similar to that of Cluster 2, but the estimates of direct and indirect effects are not.

In Web Appendix G, we fit two simple models (II) and (III) which are specified in Section 5 to compare their estimates and clusters with the results from the Bayesian profile regression model. Since the two models each require a pre-determined number of clusters, we specify $K = 3$ based on our findings. Unlike the Bayesian profile regression model, models (II) and (III) provide similar characteristics of the effect modifiers across the clusters. Their effect estimates also look different from the causal effect estimates from the proposed approach.

In addition to the cluster-specific effects from the best partition, our method allows estimation of the effects for any pre-specified covariate profiles. To do this, we can calculate assignment probability conditioning on a pre-specified profile, $\mathbf{o} = \{o_1, \dots, o_P\}$, for each MCMC iteration and calculate weighted averages of $\alpha_{1,k}^{(m)}, \alpha_{3,k}^{(m)}, \beta_{0,k}^{(m)}, \mathbf{f}_k^{(m)}$ with respect to the assignment probability of the pre-specified profile. The effects of the pre-specified profile are functions of the causal estimates described in Section 3.1. The assignment probability of the pre-specified profile, $\mathbf{o} = \{o_1, \dots, o_P\}$, to the k -th cluster can be estimated using

$$Pr(Z=k|\mathbf{u}=\mathbf{o}) \approx Pr(Z=k) \prod_{p=1}^P Pr(u_p=o_p|Z=k) = \psi_k \prod_{p=1}^P \phi_k^p(u_p=o_p).$$

Table 5 presents the effects of two pre-specified covariate profiles. Since, the relationship between BMI and diabetes and between tobacco and diabetes and their impact via socio-economic status and other diseases have been studied extensively, we examine the extent to which the impact of the expressive writing varies by different combinations of BMI, diabetes and smoking status. The expressive writing program appears to be more effective (on average) among obese patients via both direct and indirect pathways.

7 Conclusion

We have proposed a Bayesian latent mediation model which estimates cluster-specific effects, where each cluster is determined by individual covariate profiles and individual effects due to unobserved characteristics. We minimized parametric assumptions for the observed data model by using a smooth nonlinear function for the mediator. Under sequential ignorability, we show the cluster-specific effects are causal effects. A major advantage of the proposed approach is that it estimates direct and indirect effects for

subpopulations. To the best of our knowledge, our approach is the first one to address this issue in the mediation analysis setting. **The error terms in the outcome and mediator models follow a Gaussian distribution assumption; this assumption is relaxed via cluster-specific error distributions. In the simulation study, the proposed approach was not sensitive to non-normal error terms.**

For the RCC data, we identified three clusters. On average, expressive writing had a larger impact on reducing the MDASI among those patients in the second cluster where most of the patients had not received any systemic treatment. The cluster-specific effects were not statistically significant, but this is likely due to the insufficient number of subjects in the data example.

Future work will incorporate multiple mediators in this analysis. This could be done by specifying additional models for the mediators and additional identifying assumptions (e.g., an extension of sequential ignorability). We will also explore non-ignorable missingness for the outcome and the mediator. **Another potentially useful extension is to develop models for categorical outcomes and mediators. This can be done by replacing linear models with generalized models in (2) and (3) and considering the mediator as a factor in the outcome model. However, in that case, we cannot preserve conjugacy in posterior computations and need to derive a different posterior computation algorithm (e.g., Metropolis-Hastings) which may not be as efficient as the current algorithm. Also, the cluster-specific effects will not be expressed in closed forms as in Section 3.**

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by the National Institutes of Health [R01CA183854, R01GM112327, P30CA016672], the National Cancer Institute [R01CA090966], and the Richard E. Haynes Distinguished Professorship in Clinical Cancer Prevention to L.C.

The simulations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. *Conflict of Interest:* None declared.

References

1. Pearl, J. Causality. New York, NY: Cambridge University Press; 2000.
2. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992; 3(2):143–155. [PubMed: 1576220]
3. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. 1986; 51(6):1173–1182. [PubMed: 3806354]
4. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*. 2010; 25(1):51–71.
5. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annual review of psychology*. 2007; 58:593–614.

6. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974; 66(5):688–701.
7. Rubin DB. Comment on “Randomization analysis of experimental data: The Fisher randomization test comment” by Basu D. *Journal of the American Statistical Association*. 1980; 75(371):591–593.
8. Mealli F, Rubin DB. Assumptions allowing the estimation of direct causal effects. *Journal of Econometrics*. 2003; 112(1):79–87.
9. Mealli F, Mattei A. A refreshing account of principal stratification. *The International Journal of Biostatistics*. 2012; 8(1):1–17.
10. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58(1):21–29. [PubMed: 11890317]
11. Gallop R, Small DS, Lin JY, Elliott MR, Joffe M, Ten Have TR. Mediation Analysis with Principal Stratification. *Statistics in Medicine*. 2009; 28(7):1108–1130. [PubMed: 19184975]
12. Elliott MR, Raghunathan TE, Li Y. Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*. 2010; 11(2):353–372. [PubMed: 20101045]
13. Jin H, Rubin DB. Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*. 2008; 103(481):101–111.
14. Schwartz SL, Li F, Mealli F. A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*. 2011; 106(496):1331–1344.
15. VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters*. 2008; 78(17):2957–2962.
16. VanderWeele TJ. Principal stratification—uses and limitations. *The International Journal of Biostatistics*. 2011; 7(1):1–14.
17. Yuan Y, MacKinnon DP. Bayesian mediation analysis. *Psychological Methods*. 2009; 14(4):301–322. [PubMed: 19968395]
18. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National survey of children’s health. *Biostatistics*. 2010; 11(3):484–498. [PubMed: 20350957]
19. Molitor J, Su JG, Molitor N-T, et al. Identifying vulnerable populations through an examination of the association between multipollutant profiles and poverty. *Environmental Science & Technology*. 2011; 45(18):7754–7760. [PubMed: 21797252]
20. VanderWeele TJ. Confounding and effect modification: distribution and measure. *Epidemiologic Methods*. 2012; 1(1):55–82. [PubMed: 25473593]
21. Forastiere L, Mealli F, VanderWeele TJ. Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: Disentangling Bed Nets using Bayesian Principal Stratification. *Journal of the American Statistical Association*. 2016; 111(514):510–525. [PubMed: 28008210]
22. Muller D, Judd CM, Yzerbyt VY. When moderation is mediated and mediation is moderated. *Journal of personality and social psychology*. 2005; 89(6):852–863. [PubMed: 16393020]
23. Preacher KJ, Rucker DD, Hayes AF. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate behavioral research*. 2007; 42(1):185–227. [PubMed: 26821081]
24. Muthén B, Asparouhov T. Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*. 2015; 22(1):12–23.
25. Milbury K, Spelman A, Wood C, et al. Randomized controlled trial of expressive writing for patients with renal cell carcinoma. *Journal of Clinical Oncology*. 2014; 32(7):663–670. [PubMed: 24470003]
26. Crainiceanu CM, Ruppert D, Wand MP. Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software*. 2005; 14(14):1–24.
27. Wood SN. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2003; 65(1):95–114.
28. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*. 2001; 96(453):161–173.

29. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica*. 1994; 4(2):639–650.
30. Mattei A, Mealli F. Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2011; 73(5):729–752.
31. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass)*. 2010; 21(4):540.
32. Tchetgen Tchetgen EJ, Shpitser I. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*. 2012; 40(3):1816. [PubMed: 26770002]
33. Kim C, Daniels MJ, Marcus BH, Roy JA. A Framework for Bayesian Nonparametric Inference for Causal effects of Mediation. *Biometrics*. 2017; 73(2):401–409. [PubMed: 27479682]
34. Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM: an R package for profile regression mixture models using Dirichlet processes. 2013; 64(7) arXiv:1303.2836.
35. Stephens M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2000; 62(4):795–809.
36. Kaufman, L., Rousseeuw, PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. New Jersey: John Wiley & Sons; 2009.
37. Hastie, TJ., Tibshirani, RJ. *Generalized Additive Models*. New York: CRC press; 1990.
38. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. Mediation: R package for causal mediation analysis. *Journal of Statistical Software*. 2014; 59(5):1–38. [PubMed: 26917999]
39. Hastie DI, Liverani S, Richardson S. Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*. 2015; 25(5):1023–1037. [PubMed: 26321800]

Figure 1.
The structure of the simulated dataset.

Figure 2. Posterior means of probabilities for the selected covariates having pronounced differences across the clusters (stage, treatment and surgery).

Table 1

Data generating distributions for the simulations. $SN(s_1, s_2, s_3)$ denotes a skew-normal distribution with location parameter s_1 , scale parameter s_2 , and shape parameter s_3 .

Covariates	Cluster A	Cluster B	Cluster C
U_1	$N(0, 0.1)$	$N(1, 0.1)$	$N(1.5, 0.1)$
U_2	$N(1, 0.1)$	$N(2, 0.1)$	$N(0.5, 0.1)$
U_3	$Ca(0.1, 0.2, 0.7)$	$Ca(0.7, 0.2, 0.1)$	$Ca(0.1, 0.2, 0.7)$
U_4	$Ca(0, 0.1, 0.1, 0.8)$	$Ca(0.8, 0.1, 0.1, 0)$	$Ca(0.6, 0.2, 0.1, 0.1)$
U_5	$Ca(0, 0.1, 0.3, 0.6)$	$Ca(0.6, 0.3, 0.1, 0)$	$Ca(0.7, 0.1, 0.1, 0.1)$
U_6	$Ca(0, 0, 0.1, 0.2, 0.7)$	$Ca(0.7, 0.2, 0.1, 0, 0)$	$Ca(0, 0, 0.1, 0.1, 0.8)$
$M \& Y$	Cluster	Model	
M	Cluster a	$N(3 + 0.3X + 0.5U_1 + 1U_2, 0.5)$	
	Cluster b	$N(1 - 0.5X + 1.5U_1 + 1.5U_2, 0.5)$	
	Cluster c	$N(3 + 0.6X + 0.5U_1 + 0.5U_2, 0.5)$	
Y	Cluster a	$N(-2 - 1X + 0.5 \sin(10M) + 0.2U_1 + 0.4U_2, 3, 0.5)$	
	Cluster b	$N(-1 - 1X - 1M + 0.5U_1 + 0.5U_2, 1.5)$	
	Cluster c	$SN(1X + 2M + 0.3U_1 + 0.3U_2, 1, 3)$	

Simulation results for point estimators of NIE, NDE and TE over 500 replications. The columns correspond to bias and MSE relative to the true values of NIE, NDE and TE for each cluster under three different models; Model (I) : a standard frequentist model without clustering; Model (II) : a standard frequentist model with clustering covariates; Model (III) : a standard frequentist model with clustering the entire dataset; Model (IV) : our Bayesian profile regression model. Bold numbers indicate absolute biases larger than 0.15 or MSEs larger than 0.2.

Table 2

Truth	Cluster 1				Cluster 2				Cluster 3				Cluster 4			
	0				0.5				1.2				1.2			
	NIE	NDE	TE		NIE	NDE	TE		NIE	NDE	TE		NIE	NDE	TE	
Scenario 1	(I)	NIE	Bias (MSE)	0.94 (1.03)	0.44 (0.34)				-0.26 (0.22)				-0.26 (0.22)			
		NDE	Bias (MSE)	0.34 (0.95)	0.34 (0.95)				-1.65 (3.58)				-1.65 (3.58)			
		TE	Bias (MSE)	1.27 (2.58)	0.77 (1.54)				-1.92 (4.64)				-1.92 (4.64)			
	(II)	NIE	Bias (MSE)	0.07 (0.07)	-0.28 (1.47)				-0.44 (1.91)				-0.12 (1.10)			
		NDE	Bias (MSE)	0.21 (0.29)	1.42 (6.69)				-0.75 (5.40)				-0.11 (2.19)			
		TE	Bias (MSE)	0.27 (0.42)	1.13 (6.89)				-1.19 (7.17)				-0.22 (3.30)			
	(III)	NIE	Bias (MSE)	0.05 (0.03)	-0.15 (0.16)				-0.29 (0.34)				-0.39 (0.64)			
		NDE	Bias (MSE)	0.07 (0.05)	0.69 (1.16)				-0.10 (0.06)				-0.44 (0.78)			
		TE	Bias (MSE)	0.12 (0.11)	0.55 (1.06)				-0.39 (0.56)				-0.84 (2.29)			
	(IV)	NIE	Bias (MSE)	-0.05 (0.02)	-0.12 (0.08)				-0.03 (0.03)				0.11 (0.14)			
		NDE	Bias (MSE)	-0.01 (0.04)	0.01 (0.14)				-0.01 (0.03)				-0.02 (0.09)			
		TE	Bias (MSE)	-0.06 (0.05)	-0.10 (0.14)				-0.04 (0.05)				0.08 (0.18)			
Scenario 2	(I)	NIE	Bias (MSE)	0.67 (0.57)	-0.17 (0.15)				-0.53 (0.40)				-0.53 (0.40)			
		NDE	Bias (MSE)	1.26 (2.69)	1.26 (2.69)				-0.74 (1.65)				-0.74 (1.65)			
		TE	Bias (MSE)	1.93 (4.87)	1.43 (3.19)				-1.27 (2.75)				-1.27 (2.75)			
	(II)	NIE	Bias (MSE)	0.06 (0.07)	-0.40 (1.23)				-0.56 (1.75)				-0.19 (1.48)			
		NDE	Bias (MSE)	0.31 (0.36)	1.43 (6.08)				-0.75 (6.59)				-0.19 (3.69)			
		TE	Bias (MSE)	0.37 (0.49)	1.03 (6.15)				-1.31 (8.83)				-0.39 (5.66)			
	(III)	NIE	Bias (MSE)	0.05 (0.04)	-0.14 (0.17)				-0.32 (0.37)				-0.35 (0.62)			

Truth	Cluster 1			Cluster 2			Cluster 3			Cluster 4								
	NIE	0	0.5	NDE	-1	-1	NDE	0	0.5	NDE	-1	-1	NDE	0	0.5	NDE	-1	-1
	TE	-1	-0.5	TE	-1	-0.5	TE	-1	-0.5	TE	-1	-0.5	TE	-1	-0.5	TE	-1	-0.5
Scenario 1	(I)	NDE	Bias (MSE)	0.07 (0.06)	0.68 (1.11)	-0.11 (0.06)	-0.43 (0.72)											
		TE	Bias (MSE)	0.11 (0.11)	0.53 (0.93)	-0.43 (0.62)	-0.79 (2.12)											
		NIE	Bias (MSE)	-0.05 (0.02)	-0.13 (0.09)	-0.07 (0.16)	0.09 (0.09)											
	(II)	NDE	Bias (MSE)	0 (0.04)	0.06 (0.20)	0.01 (0.03)	0.04 (0.07)											
		TE	Bias (MSE)	-0.05 (0.05)	-0.07 (0.19)	-0.06 (0.19)	0.13 (0.18)											
		NIE	Bias (MSE)	0.94 (0.91)	0.44 (0.22)	-0.26 (0.09)	-0.26 (0.09)											
	(III)	NDE	Bias (MSE)	0.35 (0.29)	0.35 (0.29)	-1.64 (2.87)	-1.64 (2.87)											
		TE	Bias (MSE)	1.29 (1.86)	0.79 (0.81)	-1.91 (3.82)	-1.91 (3.82)											
		NIE	Bias (MSE)	0.08 (0.02)	-0.17 (0.30)	-0.66 (0.69)	-0.19 (0.45)											
	(IV)	NDE	Bias (MSE)	0.29 (0.23)	1.17 (2.27)	-0.98 (1.91)	-0.28 (0.97)											
		TE	Bias (MSE)	0.38 (0.35)	1.01 (1.98)	-1.64 (3.95)	-0.47 (1.71)											
		NIE	Bias (MSE)	0.04 (0.01)	-0.12 (0.09)	-0.31 (0.33)	-0.44 (0.47)											
Scenario 2	(I)	NDE	Bias (MSE)	0.07 (0.02)	0.75 (1.16)	-0.11 (0.04)	-0.47 (0.65)											
		TE	Bias (MSE)	0.11 (0.05)	0.63 (1.05)	-0.42 (0.59)	-0.92 (2.09)											
		NIE	Bias (MSE)	0 (0.01)	0 (0.01)	0.02 (0.01)	0.09 (0.03)											
	(II)	NDE	Bias (MSE)	0 (0.01)	0 (0.02)	0 (0.01)	0 (0.01)											
		TE	Bias (MSE)	0 (0.01)	0 (0.02)	0.02 (0.01)	0.09 (0.04)											
		NIE	Bias (MSE)	0.68 (0.49)	0.18 (0.06)	-0.51 (0.28)	-0.51 (0.28)											
	(III)	NDE	Bias (MSE)	1.16 (1.55)	1.16 (1.55)	-0.83 (0.89)	-0.83 (0.89)											
		TE	Bias (MSE)	1.84 (3.61)	1.35 (2.01)	-1.35 (2.02)	-1.35 (2.02)											
		NIE	Bias (MSE)	0.03 (0.01)	-0.24 (0.26)	-0.71 (0.75)	-0.17 (0.31)											
	(IV)	NDE	Bias (MSE)	0.29 (0.20)	1.22 (2.46)	-0.82 (1.64)	-0.21 (0.65)											
		TE	Bias (MSE)	0.31 (0.28)	0.97 (2.09)	-1.53 (3.51)	-0.38 (1.23)											
		NIE	Bias (MSE)	0.04 (0.01)	-0.09 (0.11)	-0.29 (0.29)	-0.46 (0.47)											

Truth	Cluster			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	NIE	0	0.5	1.2
	NDE	-1	-1	1
	TE	-1	-0.5	2.2
(IV)	NDE	Bias (MSE)	0.07 (0.02)	0.69 (1.08)
	TE	Bias (MSE)	0.10 (0.05)	0.59 (1.08)
	NIE	Bias (MSE)	-0.02 (0.01)	-0.03 (0.01)
	NDE	Bias (MSE)	-0.04 (0.01)	0.06 (0.02)
	TE	Bias (MSE)	-0.06 (0.01)	0.02 (0.02)
			-0.03 (0.01)	0.07 (0.04)
			-0.10 (0.04)	-0.47 (0.63)
			-0.39 (0.51)	-0.93 (2.06)

Simulation results for scenario 1 for point estimators of NIE, NDE and TE over 500 replications (n=300) using the proposed approach with 1) setting $n_s = 1.5$ and 2; or 2) setting $K = 15$.

Table 3

		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Truth	NIE	0	0.5	1.2	1.2
	NDE	1	-1	1	1
	TE	-1	-0.5	2.2	2.2
$n_s = 1.5$	NIE	Bias (MSE)	-0.05 (0.02)	-0.13 (0.09)	-0.07 (0.16)
	NDE	Bias (MSE)	0 (0.04)	0.06 (0.21)	0.01 (0.03)
	TE	Bias (MSE)	-0.05 (0.05)	-0.07 (0.19)	-0.06 (0.19)
$n_s = 2$	NIE	Bias (MSE)	-0.05 (0.02)	-0.12 (0.08)	-0.03 (0.03)
	NDE	Bias (MSE)	-0.01 (0.04)	0.01 (0.14)	-0.01 (0.03)
	TE	Bias (MSE)	-0.06 (0.05)	-0.10 (0.14)	-0.04 (0.05)
$K = 15$	NIE	Bias (MSE)	-0.03 (0.01)	-0.16 (0.36)	-0.05 (0.31)
	NDE	Bias (MSE)	0.001 (0.04)	0.06 (0.25)	-0.02 (0.08)
	TE	Bias (MSE)	-0.03 (0.04)	-0.09 (0.43)	-0.07 (0.41)

Table 4

Posterior means (and 95% credible intervals) of the indirect, direct and total effects for 3 clusters from the best partition (determined through postprocessing steps in Section 5) in the data example.

		TE	DE	IE
Cluster1	Mean (95% C.I.)	-0.19 (-0.77, 0.43)	-0.18 (-0.49, 0.13)	-0.02 (-0.54, 0.52)
Cluster2	Mean (95% C.I.)	-0.33 (-1.60, 1.34)	-0.44 (-1.17, 0.42)	0.12 (-0.92, 1.66)
Cluster3	Mean (95% C.I.)	-0.15 (-1.23, 0.69)	0.13 (-0.49, 0.71)	-0.28 (-1.27, 0.34)

Posterior means (and 95% C.I.s) of the indirect, direct and total effects for two pre-specified covariate profiles for the data example. Description for each covariate variable is in Table S.1.

Table 5

Case	BMI	Diabetes	Smoking	DE	IE	TE
(1)	Normal	No	Never	-0.17 (-0.48, 0.12)	-0.04 (-0.56, 0.57)	-0.21 (-0.77, 0.44)
(2)	Obese	No	Never	-0.21 (-0.54, 0.12)	-0.05 (-0.64, 0.61)	-0.26 (-0.90, 0.51)