Journal of Statistical Theory and Practice (2023) 17:51
https://doi.org/10.1007/s42519-023-00346-3

ORIGINAL ARTICLE

Mediation Analysis in Categorical Variables under Non-Ignorable Missing Data Mechanisms

Sarah Ayoku[1] · Haresh Rochani[1] · Hani Samawi[1] · Jingjing Yin[1]

Accepted: 29 September 2023 / Published online: 1 November 2023

Abstract
A mediating variable is a variable that is intermediate in the causal path relating an independent variable to a dependent variable in statistical analysis. The mediation analysis of using a categorical predictor, mediator, and outcome variables has been investigated in the literature. It is extremely common to have missing data even after having a well-controlled study. It is also well known that missingness, especially the non-ignorable missing, in a dataset has often been proven to produce biased results. This paper uses the extended Baker, Rosenberger, and Dersimonian (BRD) model to estimate the mediation effect under non-ignorable missing mechanisms. This paper also proposes four identifiable models to estimate the mediation effect for missingness in one categorical variable with two fully observed categorical variables. We reported the relative bias and Mean Square Error to compare the performance of the proposed BRD models against the Complete Case and Multiple Imputation methods in estimating the mediated effect ($\widehat{ab}$) under the non-ignorable missing mechanism. The application of these models in estimating the mediated effect was demonstrated using the Multiple Risk Factor Intervention Trial datasets.

Keywords Mediation analysis · Mediated effect · Log-linear · Categorical variables · Logistic regression · Contingency table · Missing data · Maximum likelihood method · BRD models

1 Introduction

Mediation analysis plays a significant role in the exploration of a causal relationship between two variables. A mediation model effect focuses on how two variables are related directly or indirectly: for example, consider the presence or absence of coronary

✉ Haresh Rochani
hrochani@georgiasouthern.edu

[1] Department of Biostatistics, Epidemiology and Environment Health Sciences, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30460, USA

heart disease (CHD) in high-risk men given as the independent variable (X), death due to myocardial infarction given as the dependent variable (Y), mediated by smoking where the smoking variable is the mediator (M).

Mediation analysis, which was first developed in the psychological sciences, is now instrumental in other disease development mechanisms for identifying intermediate factors useful for treatments and clinical trials. Methodological applications have increased over the past years, and more progress has also been made in understanding and applying mediation analysis in various research fields. Wright [1] proposed the mechanism of mediation by using "Wright's path analysis." His model demonstrates the mathematical equations and diagrammatical representation in understanding the causal relationship between two variables such that the equations included the coefficients. In contrast, the diagrammatic representation included arrows to illustrate the relation's direction. These path coefficients were useful in defining the mediation effect. Wright [2] showed that although path analysis can be useful in quantifying causal relationships, it was extremely challenging to determine a causal effect between two variables. Several studies criticize path analysis and suggest more knowledge is still needed in identifying causal relations. The use of mediation analysis for research purposes also requires more in-depth information [3].

However, the first mediation hypothetically used was in stimulus-organism-response (S–O–R) [4]. The mediation analysis concept has been applied in various areas, including psychology, medical sciences, epidemiology, and clinical trials [5]. Fisher [6] introduced the use of covariate as a third variable. Later on, Lazarsfeld and his colleague Kendall [7] worked on the expansion method to explain the relationship between the two variables to the third variable. According to Wright's path analysis model, economists and sociologists could generalize the covariance model [8–11]. Hence, this model was called the structural equation model, which improved the estimated mediated effects' accuracy. Sobel [12] used the structural equation model to reduce the direct and indirect effects of standard errors and then used the standard errors in computing the mediated effects' confidence interval. More studies and research to identify the complexity of the causal effect in mediation analysis have been made [13], [14], [3].

The simplest mediation model, which consists of one mediator, is known as the single mediator model. In this model, in addition to the direct effect, the independent variable X affects the dependent variable Y through a mediator M. The variables in a mediation model can either be continuous or categorical. The ordinary regression model is used to analyze continuous variables, while logistic regression is used to analyze categorical variables.

Missing data are a challenge affecting datasets and medical records in many areas of research. It can also occur in mediation analysis, and improperly handling this missingness may introduce biased mediation effects. The majority of statistical modeling approaches are designed for complete observations for the variables included in the data. It is crucial to deal with missing data using various methods to have valid inferences. Rubin [15] introduced the taxonomy of missing data mechanisms, widely used in the statistical literature. The methods of handling data with missing observations depend on the underlying assumption of the missing data mechanisms, which

are Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR).

Categorical mediation data for simple analyses can also be presented in terms of contingency tables. Missing counts in contingency tables are important in missing data analysis, and there exist various methods for dealing with missingness in contingency tables. We can use Model-based procedures such as the Baker, Rosenberger, and Dersimonian (BRD) Models [16]. These models were proposed by Baker, Rosenberger, and Dersimonian (BRD) for analyzing missing counts in a two-way contingency table with three supplementary margins, using log-linear and maximum likelihood estimates [16]. In two-way contingency tables, the cell counts adjustments for log-linear factorization of likelihood methods have been used in several research papers. Hocking and Oxspring [17] also explained the use of maximum likelihood estimation in factoring partially classified contingency tables. Several publications recommend using log-linear models for partially classified contingency tables using conditional probabilities [18]. This paper focuses on the application of estimation of the mediation effect under the non-ignorable missing data mechanism (MNAR) using the extension of Baker, Rosenberger, and Dersimonian (BRD) models proposed by Rochani et al. [19] for a three-way contingency table. Estimation of mediation effect by BRD model approach has two main advantages over other available methods for non-ignorable data. First, the BRD approach explicitly models the missing mechanism, which will result in a full likelihood specification of the models with unique interpretations. Second, the estimation method will not be affected by the proportion of missing information which can affect the rate of convergence of methods like the Expected maximum (EM) algorithm.

Section 2 will focus on an overview of existing methods used for mediation analysis for categorical variables under non-ignorable missing data mechanisms. New models derived using the BRD model approach used for at least two non-missing categorical variables in the mediation analysis will be discussed in Sect. 3. Simulations are presented in Sect. 4. We will include the application of proposed models using the Multiple Risk Factor Intervention Trial (MRFIT) data for the Prevention of Coronary Heart Disease in Sect. 5, followed by a discussion in Sect. 6.

2 Mediation Analysis under Non-Ignorable Missing Data Mechanisms

Several methods exist in analyzing the mediation effect of continuous and categorical variables under non-ignorable missing data mechanisms. However, the purpose of this paper focuses on the non-ignorable missing data mechanism (MNAR) using the Baker, Rosenberger, and Dersimonian (BRD) model approach in analyzing the mediation models using categorical mediation variables.

Given a mediation model, the relationship between smoking and coronary heart disease can be denoted by the regression coefficient parameter, "a." On the other hand, the regression coefficient, which explains the relationship between the presence or absence of coronary heart disease in high-risk men and death due to myocardial infraction when controlling for smoking, can be denoted as "c'," which is also called the direct effect. The coefficient parameter used to describe the relationship between

smoking and death due to myocardial infarction can be denoted as "*b.*" Given these coefficients', the product value of "*a*" and "*b*" is called the indirect effect. Both the direct effect and the indirect effect make up the total effect *c*. Figure 1 represents the path diagram and equations for the mediation model.

Researchers used mediation analysis to test the difference between the total effect *c* and direct effect c'. As a rule of thumb, a mediator is considered significant in the model if the value of *(c- c')* is greater than 20% [3]. Moreover, this is interpreted that X's independent variable affects the dependent variable Y indirectly via a mediator M. In general, for any mediation model with categorical or continuous variables, the model population coefficient *a* can be calculated as:

$$\widehat{a} = \frac{Cov[X, M]}{Var[X]}, \tag{1}$$

where $Cov[X, M]$ is the covariance between variables X and M, and Var[X] is the variance of X. The model population coefficient of b and c' are given, respectively, as:

$$\widehat{b} = \frac{Var[X]Cov[X, Y] - Cov[X, M]Cov[X, Y]}{Var[X]Var[M] - Cov[X, M]^2} \tag{2}$$

$$\widehat{c'} = \frac{Var[M]Cov[X, Y] - Cov[X, M]Cov[M, Y]}{Var[X]Var[M] - Cov[X, M]^2}, \tag{3}$$

In the mediation analysis for categorical variables, logistic regression analysis is recommended when at least the dependent variable Y is categorical. Logistic regression has become well known in numerous fields, one of which is its easy transformation to the odds ratio. The equivalence ($\hat{a}\hat{b} = \hat{c} - \hat{c'}$) is true when the dependent variable Y is continuous in calculating the mediated effect. However, this is not true when the dependent variable Y is categorical. The standard error is more complicated because the covariance between $\hat{c}$ and $\hat{c'}$ for ordinary regression does not directly apply to logistic regression. Hence, this makes the mediation effect estimation also complicated in terms of computing. Samawi et al. [20] developed a more straightforward method of analyzing the mediated effect among three variables when the dependent and mediator variables were dichotomous using a new approach called the latent variable technique to adjust for $ab = c - c'$.

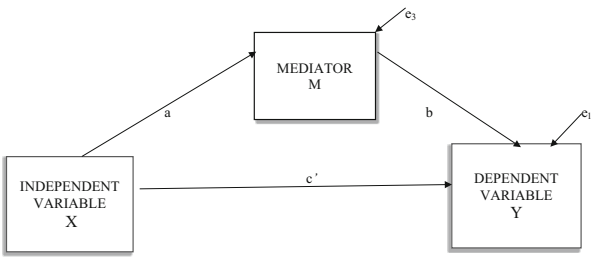

Fig. 1 Path diagram and equations for the mediation model

Although the logistic regression can analyze the categorical variables, the categorical dependent variable's scale Y^* cannot be observed directly. The residual variance and error terms are not the same as in terms of ordinary linear regression as explained by MacKinnon, and Dwyer [21]. This is because in ordinary linear regression, the dependent variance is observed and constant across the models while in logistic regression, the residual variance is fixed across the models. Winship and Mare [22] recommended setting the residual variance to $\frac{\pi^2}{3}$ to fix the scale of the unobserved dependent Y^* variable and hence the variance of Y^* becomes

$$\sigma^2_{Y^*} = \hat{c}^2\sigma^2_X + \frac{\pi^2}{3} \tag{4}$$

where Eq. (4) is the scale of the unobserved dependent Y^* for the model of the independent variable X predicting the dependent variable Y.

By applying this recommendation, MacKinnon, and Dwyer [21] showed using a simulation study that the mediation effect estimation of $\hat{c} - \hat{c'}$ and $\hat{a}\hat{b}$ were approximately equal either in the logistic or probit regression. However, in probit regression analysis, $\frac{\pi^2}{3}$ it will be replaced with 1.

There are several methods considered for modeling missingness in mediation models with categorical variables. The commonly used method for dealing with missingness in categorical data is to substitute the missing values of each observation with the most common observation value. Although this method has been proven a common approach, its challenge is that it does not consider dependencies among the observation values. Other widely used methods include complete case analysis, multiple imputation, and Model-based analysis. This paper focuses on the extension of the model-based analysis method called Baker, Rosenberger, and Dersimonian (BRD) Models to a three-way contingency table proposed by Rochani et al. [19] for estimation of the mediation effect under the non-ignorable missing data mechanism (MNAR).

3 Estimation of Models

As mentioned earlier, a contingency table can represent categorical variables for mediation analysis, especially a three-way table for a simple medication model. An illustration of a three-way contingency table with supplementary margins for analyzing the association between two binary variables *I, J* while controlling for a third variable *K* in a $2 \times 2 \times 2$ contingency table is given in Table 1.

These three-way tables with supplementary margins are used to apply log-linear models in the analysis for contingency tables with missing counts, where the missing data indicator for variable I is denoted as R_I ($R_I = 1$ represents observed values for I and $R_I = 2$ represents missing values for I). Similarly, R_J is an indicator for J's missing data such that $R_J = 1$ represents observed values for J and $R_J = 2$ represents missing values for J. The same approach applies for R_K where $R_K = 1$ indicates that K is observed. The cell counts are denoted as $\{n_{ijkab1}\}$, where *i, j*, and *k* are the levels for variables *I, J*, and *K*. The subscript *a* and *b,* when equal to 1, shows that *I* and *J* have been observed for the comparable cell and vice versa. The cell count n_{+jk211}

Table 1 Three-way table with supplementary margins

	K	I	J		
			$R_J = 1$		$R_J = 2$
			$J = 1$	$J = 2$	
$R_K = 1$	$K = 1$	$I = 1$	n_{111111}	n_{121111}	n_{1+1121}
		$I = 2$	n_{211111}	n_{221111}	n_{2+1121}
		$R_I = 1$	n_{+11211}	n_{+21211}	n_{++1221}
	$K = 2$	$I = 1$	n_{112111}	n_{122111}	n_{1+2121}
		$I = 2$	n_{212111}	n_{222111}	n_{2+2121}
		$R_I = 1$	n_{+12211}	n_{+22211}	n_{++2221}

shows where j and k and both are observed but i is missing. The cell count n_{i+k121} shows where i and k and both are observed but j is missing. Furthermore, cell count n_{++k221} shows where i and j and both are missing but k is fully observed.

Rochani et al. [19] identified sixteen BRD models using the log-linear model to an incomplete three-way table to correct for missingness in two variables with the third variable fully observed (Fig. 2).

The Fig. 2 gives a general representation of the BRD models for the I × J two-way table with three supplementary margins. In these models, α is the missing data indicator for variable I, and β is the missing data indicator for J. The first and second subscript for parameters α and β corresponds to the variables I and J, respectively. The subscript '.' indicates that the parameter is constant over the corresponding index [16]. For example, $(\alpha_{i..}, \beta_{...})$can be interpreted as missingness in a variable I depends on its own realization, while the missingness in variable J is constant across variables I and J.

This paper identifies four special case BRD models that can be used to correct for missingness in one categorical variable with the other two categorical variables fully observed. These models are model $(\alpha_{...})$, Model (α_i), model $(\alpha_{.j.})$, and Model $(\alpha_{..k})$. Model $(\alpha_{...})$ is under the ignorable missing mechanism assumption, while the other three are under the non-ignorable missing data mechanisms.

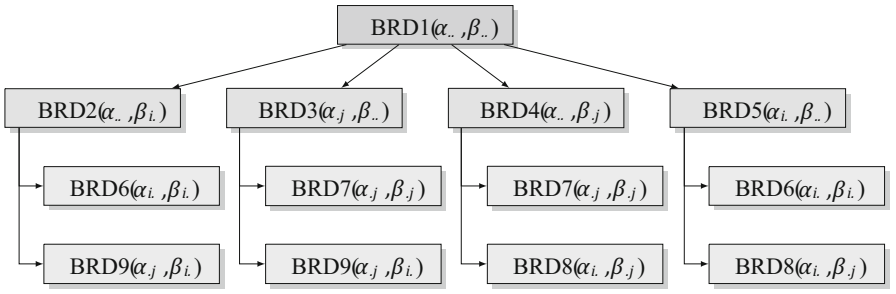

Fig. 2 Schematic presentation of BRD models [23]

In deriving the proposed models for this paper, the first procedure requires deriving the models' likelihood functions and afterward solving the system of equations for each model's maximum likelihood estimates. For illustrative purposes, we will use model $(\alpha_{i..})$ to show the parameter estimates $\widehat{\alpha}$ *and* $\widehat{M}_{ijk}$. The joint probability distribution and the log-likelihood function is given in Eq. 1 and 2, respectively.

$$L = \left\{ \prod_{i,j,k} \frac{e^{-\mu_{ijk111}} (\mu_{ijk111})^{n_{ijk111}}}{n_{ijk111}} \times \prod_{j,k} \frac{e^{-\mu_{+jk211}} (\mu_{+jk211})^{n_{+jk211}}}{n_{+jk211}} \right\} - \mu_{+++111} \tag{5}$$

$$L = \sum_i \sum_j \sum_k n_{ijk111} \log(\widehat{m}_{ijk}) + \sum_j \sum_k n_{+jk211} \log\left(\sum_i (\widehat{m}_{ijk} \widehat{\alpha}_{i..}) \right) - \sum_i \sum_j \sum_k \left\{ \widehat{m}_{ijk} (1 + \widehat{\alpha}_{i..}) \right\} + \Delta \tag{6}$$

Further simplification of the model gives the parameter estimates $\sum_i \widehat{m}_{ijk} \widehat{\alpha}_{i..} = n_{+jk211} \; \forall \, \widehat{\alpha}_{i..}$ and.

$\widehat{m}_{ijk} = n_{ijk111}$. (For a detailed derivation of these models, refer to the appendix.)

Table 2 illustrates the estimated expected cell counts using $(\alpha_{i..})$ model for complete cells and missing counts for a three-way table. Table 3 illustrates the collapsed expected cell counts for the three-way table into a $2 \times 2 \times 2$ cross-classified table, obtained by adding the cells for the estimated cell counts of the complete cells and missing cells. Based on this estimated expected count, it can be expanded into a long-form of the data and used for the analysis of fitting the mediation models and estimating the coefficients $\widehat{a}, \widehat{b}, \widehat{c}, c\prime$, and the mediation effect estimate $(\widehat{ab})$ using logistic regressions.

We can find ad hoc boundary estimates if any solution is negative, as discussed by Baker et al. [16]. Rochani et al. [19] proposed that the ML estimates can still be computed by maximizing the likelihood function using the limited memory algorithm for constrained optimization (Byrd et al., 1995).

Table 2 Estimated cell counts under model $(\alpha_{i..})$

Variable Z		Variable Y	Variable X	
	Missing		$X = 1$	$X = 2$
$Z = 1$	No	$Y = 1$	$\widehat{m}_{111}$	$\widehat{m}_{211}$
$Z = 2$		$Y = 2$	$\widehat{m}_{121}$	$\widehat{m}_{222}$
$Z = 1$	Yes	$Y = 1$	$\widehat{m}_{111}$	$\widehat{m}_{211}$
$Z = 2$		$Y = 2$	$\widehat{m}_{122}$	$\widehat{m}_{222}$
$Z = 1$		$Y = 1$	$\widehat{m}_{111}\widehat{\alpha}_1$	$\widehat{m}_{211}\widehat{\alpha}_1$
$Z = 2$		$Y = 2$	$\widehat{m}_{122}\widehat{\alpha}_2$	$\widehat{m}_{222}\widehat{\alpha}_2$

Table 3 $2 \times 2 \times 2$ cross classified table of the estimated expected counts under model $(\alpha_{i..})$

Variable Z		Variable Y	Variable X	
	Missing		$X = 1$	$X = 2$
$Z = 1$	No	$Y = 1$	$\widehat{m}_{111}$	$\widehat{m}_{211}$
$Z = 2$		$Y = 2$	$\widehat{m}_{121}$	$\widehat{m}_{222}$
$Z = 1$	Yes	$Y = 1$	$\widehat{m}_{111}(1 + \widehat{\alpha}_1)$	$\widehat{m}_{211}(1 + \widehat{\alpha}_1)$
$Z = 2$		$Y = 2$	$\widehat{m}_{122}(1 + \widehat{\alpha}_2)$	$\widehat{m}_{222}(1 + \widehat{\alpha}_2)$

4 Simulations

A simulation study was conducted to evaluate the performance of estimating the mediation effect under the non-ignorable missing mechanism by the BRD model approach compared to the complete case method and commonly used Multiple imputation method under MAR assumption. We will use the proposed model for handling missingness in one categorical variable with the other two variables are fully observed. Then under model $(\alpha_{.j.})$, the missing values were created for different percent missing in such a way that missingness in the independent variable X depends on the dependent variable Y. To model the missing probability for variable X, the following logistic regression model was considered as follows:

$$\Pr(X = \text{missing}|Y) = \frac{\exp(\gamma_0 + \beta_1 Y)}{1 + \exp(\gamma_0 + \beta_1 Y)}, \tag{7}$$

where $\beta_1 = 1$ and the choice of γ_0, which was selected by simulation, depends on the percent missing. For each iteration, sample sizes of 300, 500, and 1000 with mean $(\mu) = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ and correlation matrix $(\rho) = \begin{bmatrix} 1.000 & 0.612 & 0.125 \\ 0.612 & 1.000 & 0.612 \\ 0.125 & 0.612 & 1.000 \end{bmatrix}$ were used. This correlation matrix will give a 75% mediation effect [24]. Using a 75% mediation, the population correlations $\rho_{XM} = 0.612$, $\rho_{MY} = 0.612$ and $\rho_{XY} = 0.612$ produce $0.612 \times 0.612 = 0.3745$, which is a 3:1 ratio to 0.125. These correlations produce path coefficients of a = 0.612, b = 0.856 with the product of ab = 0.524. One thousand iterations were performed for each simulation scenario for various sample sizes and varying percentages of missingness. The multiple imputation (MI) method and the BRD model $(\alpha_{.j.})$ were used for illustrative purposes to generate expected cell counts for complete cells and missing counts. These expected counts were expanded into a long-form and used to fit the mediation models using logistic regression. Table 2 shows the biases and mean squared errors (MSEs) of the mediation effect for the complete case (CC) method, multiple imputation (MI) method, and the BRD models.

By examining the overall trend and performance in Table 4, as the percent missing in the data increases, so does the bias and MSE of the mediated effect for complete case method, model-based method, and MI method, which uses the Markov Chain Monte

Table 4 Bias and MSE comparison between complete case data, model case, and MI method data for the model (.j.)

	Relative bias (%)			MSE		
% Missing in data	CC for Mediated effect	Proposed Model for Mediated effect	MI for Mediated effect	CC for Mediated effect	Proposed Model for Mediated effect	MI for Mediated effect
$N = 300$						
10	5.7618	0.1076	9.9279	0.1664	0.1182	0.1504
20	12.6547	0.4861	30.2346	0.3983	0.1433	0.5527
30	21.0571	0.7684	50.4069	0.8041	0.1624	0.5369
40	31.5235	1.4036	76.7331	1.4183	0.1958	0.9799
50	46.7508	2.2605	103.5051	2.5061	0.2949	1.5571
$N = 500$						
10	5.6587	0.0624	13.8346	0.2061	0.1162	0.1857
20	12.4971	0.3528	41.4305	1.1407	0.1629	1.1554
30	21.0095	0.1548	72.0637	2.5399	0.1991	2.5666
40	31.2458	0.7703	106.5526	4.8966	0.2527	4.8363
50	45.7996	1.2289	149.1987	7.7914	0.3665	7.8482
$N = 1000$						
10	5.4894	0.1089	22.5950	0.5732	0.1185	0.5962
20	12.3680	0.1523	67.0049	2.1147	0.1456	2.1499
30	20.5467	0.2556	113.0949	4.8490	0.1937	4.8729
40	30.6577	0.3085	162.7302	9.2129	0.2487	4.8140
50	44.9761	0.4897	221.2242	14.8564	0.3629	7.7246

Carlo (MCMC) method for imputation. In general, the mediated effect estimates ($\widehat{ab}$) for the proposed model under the non-ignorable missing mechanism shows decreased relative biases and reduced MSE for different percent missing in the data compared to the complete case method and multiple imputation method. This shows that the application of this simulation to a simulated dataset using any of the proposed models shown in this paper will at least fit that particular simulated model as shown in this section.

5 Application to Multiple Risk Factor Intervention Trial (MRFIT) data

This section demonstrates the application of estimating the mediation effect by applying the BRD models using the Multiple Risk Factor Intervention Trial (MRFIT) data. The MRFIT dataset was available by request from the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC), which serves as the National Heart, Lung, and Blood Institute (NHLBI) biospecimens and data under

the Identifier no: NCT00000487. This dataset consisted of 12,866 men equally randomized to either an intervention or usual care group after the first two screenings. The primary endpoint in the study was death due to coronary heart disease. A total of 12,866 men were assessed to be in the upper 10–15% of CHD risk based on high serum cholesterol levels, diastolic blood pressure (BP), and cigarette use and were randomized into the study. After randomization, participants were screened annually and assessed for changes in the risk factor. The usual care group ($n = 6438$) was referred to their regular source of medical care and was examined annually. Participants in the particular intervention group ($n = 6428$) participated in an in-depth sustained multifactor intervention program to lower serum cholesterol and blood pressure and promoting smoking cessation. Participants were followed up till February 1982. Each participant was followed up for a minimum of 6 years, and the average follow-up was seven years. During follow-up, deaths were ascertained by clinical center staff, and the cause of death was determined by a committee blinded to the intervention group. The primary endpoint was CHD death and included death from MI, sudden death, Congestive heart failure (CHF), and coronary artery surgery. Other deaths from cardiovascular diseases (CVD) were from stroke, hypertension with left ventricular failure, pulmonary embolus, and unclassified CVD deaths.

Based on this MRFIT dataset, the variables of interest for this paper are the presence of coronary heart disease, which will be the independent variable, smoking which will be used as the mediating variable, and death due to myocardial infarction as the dependent variable. We will use these variables to analyze the mediation models in estimating the mediation effect using the MRFIT dataset, i.e., how smoking status mediates the relationship between coronary heart disease in high-risk men and the outcome of death due to myocardial infarction. Given the three variables of interest, missingness is present in two (Smoking and presence of coronary heart disease) while the third variable is completely observed. This is shown in Table 5.

For illustrative purposes, by focusing on one of the sixteen BRD models, say model $(\alpha_{i..}, \beta_{...})$, where α represents the missing data parameter for the smoking variable and β represents the missing data parameter for the presence of coronary heart disease variable. Hence this model implies that the participant's nonresponse on smoking depends on their smoking status and implies that the probability of missingness in

Table 5 Three-way table data representation of the MRFIT dataset with missing

		Presence of coronary heart disease		
Death due to myocardial infraction	Smoking	Yes	No	Missing
Yes	Yes	4	105	97
	No	4	89	70
	Missing	1	16	42
No	Yes	100	5031	454
	No	85	5157	502
	Missing	18	749	342

Table 6 Model comparison and parameter estimates for the MRFIT dataset

Model	Parameter a	Parameter c	Parameter c'	Parameter b	*P*-value (two-sided)	Mediation effect (If ab > 20% ofc)
$(\alpha_{...},\beta_{...})$	0.1634	− 0.8148	− 0.8049	− 0.2571	0.2893	TRUE
$(\alpha_{...},\beta_{i..})$	0.1737	− 0.8167	− 0.8091	− 0.1817	0.3896	TRUE
$(\alpha_{.j.},\beta_{...})$	0.1784	− 0.8521	− 0.8412	− 0.2616	0.2502	TRUE
$(\alpha_{...},\beta_{..k})$	0.1662	− 0.8267	− 0.8197	− 0.1762	0.3395	TRUE
$(\alpha_{..k},\beta_{...})$	0.1749	− 0.7765	− 0.7658	− 0.2595	0.2661	TRUE
$(\alpha_{...},\beta_{.j.})$	1.5365	− 2.0916	− 2.2846	0.5385	0.6808	TRUE
$(\alpha_{i..},\beta_{...})$	1.5369	− 2.3728	− 2.5371	0.4503	0.4157	TRUE
$(\alpha_{i..},\beta_{i..})$	1.5443	− 2.2151	− 2.3559	0.3887	0.3780	TRUE
$(\alpha_{.j.},\beta_{.j.})$	1.5201	− 2.0357	− 2.2014	0.4714	0.9177	TRUE
$(\alpha_{..k},\beta_{..k})$	0.1685	− 0.8281	− 0.8209	− 0.1782	0.3335	TRUE
$(\alpha_{i..},\beta_{.j.})$	1.3644	− 2.1754	− 2.3878	0.6242	0.5386	TRUE
$(\alpha_{.j.},\beta_{..k})$	0.1649	− 0.8226	− 0.8156	− 0.1769	0.3402	TRUE
$(\alpha_{..k},\beta_{i..})$	0.1685	− 0.8348	− 0.8277	− 0.1732	0.4140	TRUE
$(\alpha_{..k},\beta_{.j.})$	1.5085	− 2.0822	− 2.2627	0.5125	0.9488	TRUE
$(\alpha_{.j.},\beta_{i..})$	0.1723	− 0.1829	− 0.8166	− 0.8089	0.3900	FALSE
$(\alpha_{i..},\beta_{..k})$	2.1749	− 1.1598	− 1.0435	− 0.2884	0.8232	FALSE

the presence of coronary heart disease is independent of either presence of coronary heart disease or smoking status. This model was chosen as a more probable model under the assumption that smokers are more likely not to respond to their smoking status while missing in CHD is completely at random. However, it is always important to evaluate our conclusion's robustness by conducting a sensitivity analysis based on other non-ignorable models.

Table 6 shows the model comparison and parameter estimates for the sixteen BRD models discussed in earlier chapters using the MRFIT dataset. It is important to note that conducting a sensitivity analysis aids in the confidence of the initial assumption chosen and the conclusion made. Examining the other models will give the researcher more confidence about their hypothetical level of confidence in sticking to the initial conclusion if the conclusion does not change. Hence, from this table, there is no mediation effect based on the BRD models' p values. This implies that smoking status is not a mediating factor in the relationship between coronary heart disease in high-risk men and the outcome of death due to myocardial infarction. However, it is essential to note that in practice, decisions about having a mediation effect are often based on if the indirect effect is more than 20% of the total effect or not and not solely on a significant p-value (as shown in Table 6). Therefore, by considering this method for conclusion purposes, it is left at the discretion of the researcher to decide on which models with mediation effect is plausible for use or not.

6 Conclusion

In this paper, we have shown the application of estimation of the mediation effect under the non-ignorable missing data mechanism (MNAR) using the extension of Baker, Rosenberger, and Dersimonian (BRD) models. Generally, mediation analysis is becoming very popular in several research areas. Investigators are interested in simply knowing the cause-effect relationship between two variables; they want to understand how and why a third variable mediates this relationship. This paper illustrated how well the mediation effect's estimation under the non-ignorable missing mechanism of the BRD model approach produces accurate inference compared to either the complete case method or the MI method. Performing a sensitivity analysis based on the non-ignorable BRD models was used in evaluating the robustness of the initial assumption chosen and the conclusion made. While this paper developed a sufficient method to evaluate the performance of the estimation of the single mediation effect under the non-ignorable missing mechanism by the BRD model approach, it is recommended that further research be conducted for scenarios of multiple mediation effects. In addition, although this paper considered only categorical variables for the simple mediation model, there is also a need for future research to accommodate the mediation model variables to be a combination of continuous and categorical variables.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

Appendix

Model ($\alpha_{...}$)

$$L = \left\{ \prod_{i,j,k} \frac{e^{-\mu_{ijk111}} (\mu_{ijk111})^{n_{ijk111}}}{n_{ijk111}} \times \prod_{j,k} \frac{e^{-\mu_{+jk211}} (\mu_{+jk211})^{n_{+jk211}}}{n_{+jk211}} \right\} - \mu_{+++111},$$

given $\mu_{ijk111} = \widehat{m_{ijk}}$ and $\mu_{+jk211} = \widehat{m_{+jk}\alpha_{...}}$, the above equation can be written as

$$L = \left\{ \prod_{i,j,k} \frac{e^{-m_{ijk}} (\widehat{m}_{ijk})^{n_{ijk111}}}{n_{ijk111}} \times \prod_{j,k} \frac{e^{-\widehat{m}_{+jk}\alpha_{...}} (\widehat{m}_{+jk}\widehat{\alpha}_{...})^{n_{+jk211}}}{n_{+jk211}} \right\} - \mu_{+++111}$$

The log-likelihood function can be derived as

$$L = \sum_i \sum_j \sum_k n_{ijk111} \log(\widehat{m}_{ijk}) + \sum_j \sum_k n_{+jk211} \log(\widehat{m}_{+jk}\widehat{\alpha}_{...}) - \sum_i \sum_j \sum_k \left\{ \widehat{m}_{ijk}(1 + \widehat{\alpha}_{...}) \right\} + \Delta$$

By differentiating with respect to $\widehat{\alpha_{...}}$, we have:

$$\therefore \frac{dL}{d\widehat{\alpha}_{...}} = \frac{n_{+jk211}}{\widehat{m}_{+jk}\widehat{\alpha}_{...}} \bullet \widehat{m}_{+jk} - \widehat{m}_{+++}$$

$$\therefore 0 = \frac{n_{+jk211}}{\widehat{\alpha}_{...}} - \widehat{m}_{+++}$$

$$\therefore \widehat{\alpha}_{...} = \frac{n_{+jk211}}{\widehat{m}_{+++}}.$$

Given each cell count denoted by $\{n_{ijkabc}\}$, where $i, j,$ and k represent the categories for variables I, J, and K, respectively. Hence, we have

$$\widehat{m}_{ijk} = \widehat{m}_{+++} = n_{+++111}$$

$$n_{+jk211} = n_{+++211}$$

$$\therefore \boxed{\widehat{\alpha}_{...} = \frac{n_{+++211}}{n_{+++111}}}$$

By differentiating with respect to $\widehat{m_{+jk}}$, we have:

$$\frac{dL}{d\widehat{m}_{ijk}} = \frac{n_{ijk111}}{\widehat{m}_{ijk...}} + \frac{n_{+jk211}}{\widehat{m}_{+jk}\widehat{\alpha}_{...}} \bullet \widehat{\alpha}_{...} - (1 + \widehat{\alpha}_{...})$$

$$0 = \frac{n_{ijk111}}{\widehat{m}_{ijk...}} + \frac{n_{+jk211}}{\widehat{m}_{+jk}} - (1 + \widehat{\alpha}_{...})$$

$$\therefore (1 + \widehat{\alpha}_{...}) = \frac{n_{ijk111}}{\widehat{m}_{ijk...}} + \frac{n_{+jk211}}{\widehat{m}_{+jk}}$$

$$\therefore \left(1 + \frac{n_{+++211}}{n_{+++111}}\right) = \frac{n_{ijk111}}{\widehat{m}_{ijk...}} + \frac{n_{+jk211}}{\widehat{m}_{+jk}}$$

$$\left(\frac{n_{+++111} + n_{+++211}}{n_{+++111}}\right) = \frac{n_{ijk111}}{\widehat{m}_{ijk...}} + \frac{n_{+jk211}}{\widehat{m}_{+jk}}$$

$$(\frac{n_{++++11}}{n_{+++111}})\widehat{m}_{ijk...} = n_{ijk111} + \frac{n_{+jk211}}{\widehat{m}_{+jk}}\widehat{m}_{ijk}$$

$$\widehat{m}_{ijk...} = \left(n_{ijk111} + \frac{n_{+jk211}}{\widehat{m}_{+jk}}\widehat{m}_{ijk}\right)\left(\frac{n_{+++111}}{n_{++++11}}\right)$$

$$\hat{m_{+jk}} = \left(n_{+jk111} + \frac{n_{+jk211}}{\hat{m_{+jk}}}\hat{m_{+jk}}\right)\left(\frac{n_{+++111}}{n_{++++11}}\right)$$

$$\widehat{m}_{+jk} = \left(n_{+jk111} + n_{+jk211}\right)\left(\frac{n_{+++111}}{n_{++++11}}\right)$$

$$\boxed{\widehat{m}_{+jk} = \frac{n_{+jk+11}n_{+++111}}{n_{++++11}}}$$

This can be simplified further as

$$\frac{n_{+jk+11}n_{+++111}}{n_{++++11}} = \left(n_{+jk111} + \frac{n_{++++11}n_{+jk211}\widehat{m}_{ijk}}{n_{+jk+11}n_{+++111}}\right)\left(\frac{n_{+++111}}{n_{++++11}}\right)$$

$$\frac{n_{+jk+11}n_{+++111}}{n_{++++11}} = \left(\frac{n_{+jk111}n_{+jk+11}n_{+++111} + n_{++++11}n_{+jk211}\widehat{m}_{ijk}}{n_{+jk+11}n_{+++111}}\right)\left(\frac{n_{+++111}}{n_{++++11}}\right)$$

$$\frac{n_{+jk+11}n_{+++111}}{n_{++++11}} = \left(\frac{n_{+++111}n_{+jk111}n_{+jk+11}n_{+++111} + n_{+++111}n_{++++11}n_{+jk211}\widehat{m}_{ijk}}{n_{++++11}n_{+jk+11}n_{+++111}}\right)$$

$$n_{+jk+11}n_{+++111}n_{++++11}n_{+jk+11}n_{+++111} = n_{++++11}n_{+++111}n_{+jk111}n_{+jk+11}n_{+++111} + n_{++++11}n_{+++111}n_{++++11}n_{+jk211}\widehat{m}_{ijk}$$

$$\frac{n_{+jk+11}n_{+++111}n_{++++11}n_{+jk+11}n_{+++111} - n_{++++11}n_{+++111}n_{+jk111}n_{+jk+11}n_{+++111}}{n_{++++11}n_{+++111}n_{++++11}n_{+jk211}} = \widehat{m}_{ijk}$$

$$\frac{n_{+jk+11}n_{+++111}n_{++++11}n_{+jk+11}n_{+++111} - n_{++++11}n_{+++111}n_{+jk111}n_{+jk+11}n_{+++111}}{n_{++++11}n_{+++111}n_{++++11}n_{+jk211}} = \widehat{m}_{ijk}$$

Therefore, we have

$$\boxed{\widehat{m}_{ijk} = \frac{n_{+jk+11}n_{+++111}(n_{+jk+11} - n_{+jk111})}{n_{++++11}n_{+jk211}}}$$

Model ($\alpha_{i..}$)

$$L = \left\{\prod_{i,j,k} \frac{e^{-\mu_{ijk111}}(\mu_{ijk111})^{n_{ijk111}}}{n_{ijk111}} \times \prod_{j,k} \frac{e^{-\mu_{+jk211}}(\mu_{+jk211})^{n_{+jk211}}}{n_{+jk211}}\right\} - \mu_{+++111}$$

$$L = \left\{\prod_{i,j,k} \frac{e^{-m_{ijk}}(\widehat{m}_{ijk})^{n_{ijk111}}}{n_{ijk111}} \times \prod_{j,k} \frac{e^{-\widehat{m}_{+jk}\alpha_{\cdots}}(\widehat{m}_{ijk}\widehat{\alpha}_{i..})^{n_{+jk211}}}{n_{+jk211}}\right\} - \mu_{+++111}$$

$$L = \sum_i \sum_j \sum_k n_{ijk111} \log(\widehat{m}_{ijk}) + \sum_j \sum_k n_{+jk211} \log\left(\sum_i (\widehat{m}_{ijk}\widehat{\alpha}_{i..})\right) - \sum_i \sum_j \sum_k \left\{\widehat{m}_{ijk}(1+\widehat{\alpha}_{i..})\right\} + \Delta$$

$$\therefore \frac{dL}{d\widehat{\alpha}_{i..}} = \sum_j \sum_k \left(\frac{n_{+jk211}}{\sum_i (\widehat{m}_{ijk}\widehat{\alpha}_{i..})}\widehat{m}_{ijk}\right) - \sum_j \sum_k \widehat{m}_{ijk}$$

$$\sum_j \sum_k \widehat{m}_{ijk} = \sum_j \sum_k \left(\frac{n_{+jk211}}{\sum_i (\widehat{m}_{ijk}\widehat{\alpha}_{i..})}\widehat{m}_{ijk}\right),$$

therefore, we can deduce that given $\boxed{\sum_i \widehat{m}_{ijk}\widehat{\alpha}_{i..} = n_{+jk211}}$ $\forall\widehat{\alpha}_{i..}$, then

$$\frac{dL}{d\widehat{m}_{ijk}} = \frac{n_{ijk111}}{\widehat{m}_{ijk...}} + \frac{n_{+jk211}}{\sum_i (\widehat{m}_{ijk}\widehat{\alpha}_{i..})} \bullet \widehat{\alpha}_{i..} - (1+\widehat{\alpha}_{i..})$$

$$(1+\widehat{\alpha}_{i..}) = \frac{n_{ijk111}}{\widehat{m}_{ijk...}} + \frac{n_{+jk211}}{\sum_i (\widehat{m}_{ijk}\widehat{\alpha}_{i..})} \bullet \widehat{\alpha}_{i..}$$

$$(1+\widehat{\alpha}_{i..})\widehat{m}_{ijk...} = n_{ijk111} + \frac{n_{+jk211}}{\sum_i (\widehat{m}_{ijk}\widehat{\alpha}_{i..})}(\widehat{\alpha}_{i..}\widehat{m}_{ijk...}).$$

By assuming $\sum_i (\widehat{m}_{ijk}\widehat{\alpha}_{i..}) = n_{+jk211}$, we have:

$$(1+\widehat{\alpha}_{i..})\widehat{m}_{ijk...} = n_{ijk111} + \frac{n_{+jk211}}{n_{+jk211}}(\widehat{\alpha}_{i..}\widehat{m}_{ijk...})$$

$$(1+\widehat{\alpha}_{i..})\widehat{m}_{ijk...} = n_{ijk111} + \widehat{\alpha}_{i..}\widehat{m}_{ijk...}$$

$$\boxed{\widehat{m}ijk_{ijk...} = n_{ijk111}}$$

Model ($\alpha_{.j.}$)

$$L = \sum_i \sum_j \sum_k n_{ijk111} \log(\widehat{m}_{ijk}) + \sum_j \sum_k n_{+jk211} \log(\widehat{m}_{+jk}\widehat{\alpha}_{.j.}) - \sum_j \widehat{m}_{+j+}(1+\widehat{\alpha}_{.j.}) + \Delta$$

$$\therefore \frac{dL}{d\widehat{\alpha}_{.j.}} = \sum_k \left(\frac{n_{+jk211}}{\widehat{m}_{+jk}\widehat{\alpha}_{.j.}} \widehat{m}_{+jk} \right) - \widehat{m}_{+j+}$$

$$\therefore 0 = \frac{n_{+j+211}}{\widehat{\alpha}_{.j.}} - \widehat{m}_{+j+}$$

$$\therefore \widehat{m}_{+j+} = \frac{n_{+j+211}}{\widehat{\alpha}_{.j.}}$$

$$\boxed{\therefore \widehat{\alpha}_{.j.} = \frac{n_{+j+211}}{\widehat{m}_{+j+}}}$$

$\therefore \frac{dL}{d\widehat{m}_{ijk}} = \frac{n_{ijk111}}{\widehat{m}_{ijk}} + \frac{n_{+jk211}}{\widehat{m}_{+jk}\widehat{\alpha}_{.j.}} \widehat{\alpha}_{.j.} - (1 + \widehat{\alpha}_{.j.})$, since $\widehat{m}_{+jk}\widehat{\alpha}_{.j.} = n_{+jk211}$, then:

$$\therefore (1 + \widehat{\alpha}_{.j.}) = \frac{n_{ijk111}}{\widehat{m}_{ijk}} + \widehat{\alpha}_{.j.}$$

$$\widehat{m}_{ijk}(1 + \widehat{\alpha}_{.j.}) = n_{ijk111} + \widehat{\alpha}_{.j.}\widehat{m}_{ijk}$$

$$\boxed{\therefore \widehat{m}_{ijk} = n_{ijk111}}$$

Model ($\alpha_{..k}$)

$$L = \sum_i \sum_j \sum_k n_{ijk111} \log(\widehat{m}_{ijk}) + \sum_j \sum_k n_{+jk211} \log(\widehat{m}_{+jk}\widehat{\alpha}_{..k}) - \sum_k \widehat{m}_{++k}(1 + \widehat{\alpha}_{..k}) + \Delta$$

$$\therefore \frac{dL}{d\widehat{\alpha}_{..k}} = \sum_j \left(\frac{n_{+jk211}}{\widehat{m}_{+jk}\widehat{\alpha}_{..k}} \widehat{m}_{+jk} \right) - \widehat{m}_{++k}$$

$$\therefore 0 = \frac{n_{++k211}}{\widehat{\alpha}_{..k}} - \widehat{m}_{++k}$$

$$\therefore \widehat{m}_{++k} = \frac{n_{++k211}}{\widehat{\alpha}_{..k}}$$

$$\boxed{\therefore \widehat{\alpha}_{..k} = \frac{n_{++k211}}{\widehat{m}_{++k}}}$$

$$\therefore \frac{dL}{d\widehat{m}_{ijk}} = \frac{n_{ijk111}}{\widehat{m}_{ijk}} + \frac{n_{+jk211}}{\widehat{m}_{+jk}\widehat{\alpha}_{..k}}\widehat{\alpha}_{..k} - (1+\widehat{\alpha}_{..k}), \text{ since } \widehat{m}_{+jk}\widehat{\alpha}_{..k} = n_{+jk211}, \text{ then:}$$

$$\therefore (1+\widehat{\alpha}_{..k}) = \frac{n_{ijk111}}{\widehat{m}_{ijk}} + \widehat{\alpha}_{..k}$$

$$\widehat{m}_{ijk}(1+\widehat{\alpha}_{..k}) = n_{ijk111} + \widehat{\alpha}_{..k}\widehat{m}_{ijk}$$

$$\boxed{\therefore \widehat{m}_{ijk} = n_{ijk111}}$$

References

1. Wright S (1920) The relative importance of heredity and environment in determining the piebald pattern of Guinea pigs. Proc Natl Acad Sci 6:320–332
2. Wright S (1923) The theory of path coefficients: a reply to Niles's criticism. Genetics 8:239–255
3. MacKinnon DP (2013) Introduction to statistical mediation analysis. Routledge, New York
4. Woodworth RS (1928) Dynamic psychology. In: C. Murchison (Ed.). Psychologies of 1925 (pp 111–126). Worcester, MA; Clark University Press
5. Judd CM, Kenny DA (1981) Process analysis: estimating mediation in treatment evaluations. Eval Rev 5:602–619
6. Fisher RA (1934) Statistical methods for research workers, 5th edn. Oliver and Boyd Lt, Edinburgh, Scotland
7. Kendall PL, Lazarsfeld PF (1950) Problems of survey analysis. In: Merton RK, Lazarsfeld PF (eds) Continuities is social research: studies in the scope and method of The American soldier. Free Press, Glencoe, IL, pp 133–196
8. Jöreskog KG (1970) A general method for analysis of covariance structures. Biometrika 57:239–251
9. Jöreskog KG (1973) A general method for estimating a linear structural equation system. In: Goldberger AS, Duncan OD (eds) Structural equation models in the social sciences. Seminar Press, New York, pp 85–112
10. Keesling JW (1972) Maximum likelihood approaches to causal analysis [Unpublished doctoral dissertation]. University of Chicago
11. Wiley DE (1973) The identification problem for structural equation models with unmeasured variables. In: Goldberger AS, Duncan OD (eds) Structural equation models in the social sciences. Seminar Press, New York, pp 69–83
12. Sobel ME (1982) Asymptotic confidence intervals for indirect effects in structural equation models. Sociol Methodol 13:290–312. https://doi.org/10.2307/270723
13. Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58:21–29
14. Little R, Rubin D (2002) Statistical analysis with missing data (2nd ed.). In: Models for partially classified contingency tables, ignoring the missing-data mechanism. Wiley-Interscience
15. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
16. Baker SG, Rosenberger WF, DerSimonian R (1992) Closed-form estimates for missing counts in two-way contingency tables. Stat Med 11:643–657
17. Hocking RR, Oxspring HH (1974) The analysis of partially categorized contingency data. Biometrics 30(3):469–483
18. Bishop YMM, Fienberg SE, Holland PW (2007) Discrete multivariate analysis. Springer, New York
19. Rochani HD, Vogel RL, Samawi HM, Linder DF (2017) Estimates for cell counts and common odds ratio in three-way contingency tables by homogeneous log-linear models with missing data. Adv Stat Anal AStA 101:51–65
20. Samawi H, Cai J, Linder DF, Rochani H, Yin J (2018) A simpler approach for mediation analysis for dichotomous mediators in logistic regression. J Stat Comput Simul 88(6):1211–1227. https://doi.org/10.1080/00949655.2018.1426762

21. MacKinnon DP, Dwyer JH (1993) Estimating mediated effects in prevention studies. Eval Rev 17:144–158
22. Winship C, Mare RD (1983) Structural equations and path analysis for discrete data. Am J Sociol 89:54–110
23. Jansen I (2005) Flexible model strategies and sensitivity analysis tools for non-monotone incomplete categorical data. LUC
24. Iacobucci D (2012) Mediation analysis with categorical variables: the final frontier. J Consum Psychol 22(4):582–594