Received: 13 September 2021 | Revised: 30 August 2022 | Accepted: 22 September 2022

DOI: 10.1002/bimj.202100284

Biometrical Journal

RESEARCH ARTICLE

A latent class model to multiply impute missing treatment indicators in observational studies when inferences of the treatment effect are made using propensity score matching

Robin Mitra

Department of Statistical Science,
University College London, London
WC1E 6BT, UK

Correspondence
Robin Mitra, School of Mathematics,
Cardiff University, Cardiff, CF24 4AG,
UK.
Email: ucakrmi@ucl.ac.uk

RR Reproducible Research

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Analysts often estimate treatment effects in observational studies using propensity score matching techniques. When there are missing covariate values, analysts can multiply impute the missing data to create m completed data sets. Analysts can then estimate propensity scores on each of the completed data sets, and use these to estimate treatment effects. However, there has been relatively little attention on developing imputation models to deal with the additional problem of missing treatment indicators, perhaps due to the consequences of generating implausible imputations. However, simply ignoring the missing treatment values, akin to a complete case analysis, could also lead to problems when estimating treatment effects. We propose a latent class model to multiply impute missing treatment indicators. We illustrate its performance through simulations and with data taken from a study on determinants of children's cognitive development. This approach is seen to obtain treatment effect estimates closer to the true treatment effect than when employing conventional imputation procedures as well as compared to a complete case analysis.

KEYWORDS

latent class, missing data, multiple imputation, observational studies, propensity score

1 | INTRODUCTION

In observational studies, often the covariate distributions in the treatment and control groups are imbalanced. Analysts can use propensity score matching to reduce this imbalance (with respect to the measured covariates) and estimate treatment effects (Park et al., 2007; Rosenbaum & Rubin, 1983, 1985). In this paper, we restrict attention to a binary treatment variable. The propensity score for a particular unit, $e(\boldsymbol{x}_i)$, is defined to be the probability that a generic unit is assigned to treatment given its vector of observed covariates $\boldsymbol{x}_i$; that is, $e(\boldsymbol{x}_i) = P(T_i = 1|\boldsymbol{x}_i)$, where $T_i = 1$ if subject i receives treatment and $T_i = 0$ otherwise. Rosenbaum and Rubin (1983) show that when two large groups have the same distributions of propensity scores, the groups will have the same covariate distributions. Analysts can then select control units whose propensity scores are similar to the treated units' propensity scores, creating a matched control group whose covariates are similar in distribution to the treated group. Analysts can then estimate a treatment effect, $\hat{\tau}$, using the treated and matched control groups. Specifically, we assume that analysts estimate $\hat{\tau}$ with $\bar{Y}_T - \bar{Y}_{mc}$, where $\bar{Y}_T$ and $\bar{Y}_{mc}$ comprise the mean of the responses for the treatment and matched control units, respectively. This treatment effect will now not suffer from the bias due to imbalanced covariate distributions for those covariates contained in $\boldsymbol{x}$.

Apart from matching, there are various alternative strategies to estimate causal effects with propensity scores, these include subclassification (Hullsiek & Louis, 2002; Rosenbaum & Rubin, 1984) and propensity score weighted estimation (Lunceford & Davidian, 2004). See Williamson et al. (2012b) for a comprehensive review and illustration of different approaches to causal inference using propensity scores. While we focus on estimating treatment effects using propensity score matching, these alternative strategies could also be accommodated within the framework we propose. Nevertheless, propensity score matching is very commonly used to estimate treatment effects (Austin, 2008, 2009; Cho et al., 2007; da Veiga & Wilder, 2008; Leon et al., 2012; Nguyen, 2012). We focus on treatment effects estimated using nearest neighbor matching without replacement, although other optimal matching schemes proposed in the literature such as full matching (Rosenbaum, 1991; Stuart & Green, 2008) could also be considered. A review of matching methods for causal inferences is provided in Stuart (2010). In practice, propensity scores are unknown in observational studies and are typically estimated using a regression of T on functions of $\boldsymbol{x}$ (Hanley & Dendukuri, 2009; Setoguchi et al., 2008; Woo et al., 2008; Westreich et al., 2010). In this paper, we estimate propensity scores using logistic regression models.

When dealing with missing values arising in observational studies, some papers assume that the treatment variable is fully observed, for example, Mitra and Reiter (2011) and Qu and Lipkovich (2009), and focus on missing values in other parts of the data. However, while this may be a reasonable assumption to make in randomized controlled trials or clinical trials in general, where treatment is under the control of the investigator, this may not always be an appropriate assumption in other types of studies. The simplest strategy is to discard units with a missing treatment value, akin to a complete case analysis. However, there is the potential that this could exacerbate imbalances in the covariate distributions and bias inferences as a result (Demissie et al., 2003).

There are naturally concerns about the consequences of handling the missing treatment data improperly. Simply ignoring the problem is not an adequate solution. Molinari (2010) derives bounds on the treatment effect for different types of missing data mechanisms. This was developed further by Mebane Jr and Poast (2013), who take a Bayesian approach to incorporate uncertainty in the bounds and assess sensitivity to assumptions. However, the focus here is not specifically on developing a method to estimate treatment effects. Kennedy (2020) provide nonparametric efficiency bounds for treatment effects and use this to construct nonparametric treatment effect estimators. However, the results in Kennedy (2020) are purely theoretical and asymptotic. It is not clear how these estimators will perform when applied to real data as well as how convenient these estimators will be to implement in practice. Shortreed and Forbes (2010) empirically reviewed various standard approaches to dealing with missing treatment data in a longitudinal observational study. However, the paper addresses more complex issues such as time-dependent confounders, using a marginal structural model to estimate treatment effects rather than propensity scores. These considerations fall outside the scope of this paper. We also note the work that has been done to deal with missing values, including missing treatment values, within case control studies, such as the work of Ahn et al. (2011).

Within observational studies, the majority of the methodological research to deal with this problem has been from an inverse probability weighting perspective. Williamson et al. (2012a) construct an augmented inverse probability weighted estimator that is doubly robust for the treatment effect, but can only deal with missingness in one variable. This has been built on by Zhang et al. (2016), who develop a triply robust estimator for the treatment effect. While inverse probability weighted methods have some nice theoretical properties, these typically rely on asymptotic results, and there are particular challenges in developing measures of uncertainty. In addition, constructing relevant estimators may require some effort on the part of the analyst depending on the specific models and the estimating procedure used. Further, in these approaches, the treatment of the missing values and analysis of interest are intrinsically tied together meaning any further analysis of the data is not possible or greatly complicated. This is what motivates consideration of multiple imputation in general, which has some distinct advantages in this regard and are elaborated on below. This is the research area this paper intends to explore.

In multiple imputation, missing values are repeatedly imputed by sampling from predictive distributions conditional on the observed covariate data. A key advantage of multiple imputation over inverse propensity weighting is that treatment of the missing data and subsequent data analysis are done in two distinct stages. This allows an imputer, who is well trained in dealing with missing values, to deal with the burden of the missing data, freeing the analyst to analyze the data in largely the same way as if there were no missing values to deal with. This ties into a further advantage of multiple imputation, again not present with inverse probability weighting approaches, in that the analyst can also easily pursue further modeling, such as subdomain comparisons or regression adjustment to reduce residual imbalances (Hill, 2004; Hill et al., 2004).

When it comes to the issue of missing treatment variables, multiple imputation can also have advantages over inverse probability weighting methods when the treatment variable is derived from several others. For example, in Zhang et al. (2016), the treatment variable is body mass index (BMI), which is derived from height and weight. By restricting attention to BMI as a single incomplete variable, as done in Zhang et al. (2016), less information in the data is being exploited to learn about the missing treatment variable. Multiple imputation does not suffer from this restriction, and can, for example, impute missing BMI from models for height and weight using these variables as predictors for one another, where appropriate. The benefits and popularity of multiple imputation are acknowledged in Lee et al. (2021), who use this to deal with a large amount of missing data in their treatment variable (smoking status at 14 years of age). In particular, they recognize the potential for auxiliary variables to be included in the imputation model but not necessarily in the analysis model. The benefit of multiply imputing the treatment variable has also been noted in Sartori et al. (2005) where treatment was viewed in the context of an exposure dose. We note that multiple imputation has wider applications than just to deal with missing treatment values, being used to handle potentially misclassified treatment values due to proxy reporting when treatment effects are estimated using propensity scores (Shardell & Hicks, 2014).

Multiple imputation via chained equations (MICE) (van Buuren et al., 1999; Van Buuren, 2007) is a convenient and popular technique that can impute missing values in data sets where variables are measured on different scales, such as binary, nominal, and continuous variables. Typically a fairly moderate number of imputations is deemed to be sufficient, and convergence when using the approach has been considered in the literature (White et al., 2011; Zhu & Raghunathan, 2015). MICE could thus be used to impute the missing treatment indicators and indeed the chained equations imputation approach was used in Lee et al. (2021) for this purpose. However, there is the risk that an imputation method could generate implausible values for the treatment indicators that result in biases in the treatment effect. The importance of specifying an appropriate imputation model is noted in Breger et al. (2020) who considered imputation to deal with missing treatment values when using the G-computation algorithm to estimate treatment effects. In the investigation by Shortreed and Forbes (2010), they found that a standard imputation procedure using a logistic model did not perform well. In the context of this paper, where we estimate treatment effects using propensity score matching, we would be keen to focus our attention on the imputation of the missing treatment indicators for units lying in the treated units' covariate space. A standard imputation model might tend to impute the majority of these indicators to be treated, which could cause imbalances in the covariate distributions of the two groups and affect the resulting treatment effect estimate. In general, it is important to be able to generate plausible imputations for the missing treatment indicators taking into account the fact that the covariate spaces of the treated and control groups naturally overlap.

We propose a novel approach to impute missing treatment indicators based on a latent class model originally proposed by Mitra and Reiter (2011). The latent class model was developed to impute missing covariates in observational studies that is robust to model misspecification. The model assumes that each unit's covariates either belongs to a class corresponding to the treated units' covariate distribution, or to a class corresponding to some other covariate distribution. In doing so there is an assumption that control units can have covariates that lie in the treated units' covariate space. This formulation permits the possibility for developing an imputation model that addresses the key objective of imputing missing treatment indicators using the available information in the data and not suffer from the possibility of implausible imputations generated using standard imputation models. We also develop methods to estimate treatment effects using propensity score matching in this setting. These methods are based on those proposed in Hill (2004) and Mitra and Reiter (2016), but need to take into account the variability in the imputed treatment indicators over the imputations.

The remainder of the paper is organized as follows. In Section 2, we describe strategies to address the problem of missing treatment indicators, including the proposed latent class imputation model. In Section 3, we illustrate the performance of the latent class imputation model, and other approaches, in a simulation study. In Section 4, we illustrate the performance of the model to impute missing treatment indicators in an observational study that investigated determinants of children's cognitive development. Finally, in Section 5, we end with some concluding remarks.

2 | DEALING WITH MISSING TREATMENT INDICATORS IN OBSERVATIONAL STUDIES

In this section, we describe various strategies for dealing with missing treatment indicators. These comprise: (1) a complete case analysis, (2) standard multiple imputation using MICE, and (3) the proposed novel Bayesian imputation approach using a latent class model.

Let $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})'$, $i = 1, \dots, n$ denote a vector of p covariates for the ith unit in the study. Each unit also has a binary treatment indicator T_i where $T_i = 1$ indicates a treatment unit while $T_i = 0$ indicates a control unit. We can then define an $n \times p$ covariate matrix $\boldsymbol{X}$ where the ith row of $\boldsymbol{X}$ corresponds to $\boldsymbol{x}_i$. We can also define a vector of the n treatment indicators by $\boldsymbol{T} = (T_1, \dots, T_n)'$. As we assume missing data could be present, we also define a corresponding missing data indicator vector for $\boldsymbol{x}_i$ by $\boldsymbol{m}_i$, where $m_{ij} = 1$ indicates x_{ij} is missing and $m_{ij} = 0$ indicates x_{ij} is observed. We also define a missing indicator for the treatment variable with $m_i^T = 1$ indicating T_i is missing and $m_i^T = 0$ indicating T_i is observed. Denote $\boldsymbol{X}_{obs} = \{x_{ij} : m_{ij} = 0\}$ and $\boldsymbol{X}_{mis} = \{x_{ij} : m_{ij} = 1\}$ as the observed and missing parts of the covariate data, respectively. Likewise, define $\boldsymbol{T}_{obs} = \{T_i : m_i^T = 0\}$ and $\boldsymbol{T}_{mis} = \{T_i : m_i^T = 1\}$ as the observed and missing parts of the treatment indicator variable. Each unit i also has an outcome value denoted by Y_i. We assume that the outcome variable is fully observed.

2.1 | Complete case analysis

In a complete case analysis, we only consider those units with fully observed covariates and observed treatment indicator, that is, consider only units i where $\prod_{j=1}^{p}(1 - m_{ij})(1 - m_i^T) = 1$. With this approach, standard complete data methods could be used to estimate the treatment effect. As described in Section 1, we assume analysts estimate propensity scores by fitting a logistic regression model to the complete case data and estimate a treatment effect using these scores and nearest neighbor matching without replacement. Denote the treatment effect estimated from the complete case data using propensity score matching without replacement by $\hat{\tau}^{cc}$.

The advantage with complete case analysis is its simplicity, as it allows standard complete data methods to be applied. However, the approach can result in a greatly reduced sample size upon which to base estimates, which potentially translates into a substantial loss in information with increased variability in estimates. In the case where matching is used, discarding potentially relevant matches from the control group due to missing values being present could lead to a suboptimal matched control group being selected and thus lead to the potential for bias to be present.

2.2 | Multiple imputation

Multiple imputation is a popular approach to deal with missing values. The approach proceeds by imputing the missing values using an appropriate distribution, typically the posterior predictive distribution of the missing data conditional on the observed data. That is, we draw values from the distribution,

$$p(\boldsymbol{X}_{mis}, \boldsymbol{T}_{mis} | \boldsymbol{X}_{obs}, \boldsymbol{T}_{obs}) = \int_{\theta} p(\boldsymbol{X}_{mis}, \boldsymbol{T}_{mis} | \theta, \boldsymbol{X}_{obs}, \boldsymbol{T}_{obs}) p(\theta | \boldsymbol{X}_{obs}, \boldsymbol{T}_{obs}) d\theta,$$

where θ represents model parameters for the complete data $(\boldsymbol{X}, \boldsymbol{T})$.

In practice obtaining a closed-form expression for $p(\theta | \boldsymbol{X}_{obs}, \boldsymbol{T}_{obs})$ is typically not possible and so imputations are drawn using iterative simulation techniques such as Markov chain Monte Carlo (MCMC). This is the approach taken with our proposed latent class model in the next section. However, as multiple imputation is typically not viewed to be a computationally heavy approach, and itself relies on approximations to the posterior distribution of the quantity of interest, we also consider using the MICE approach to multiply impute missing values (van Buuren et al., 1999). This approach proceeds by imputing missing values in a sequential variable by variable manner, imputing missing values in a variable using a regression model conditional on all other variables in the data. If covariates in this regression model also have missing data, the most recent imputed values in these variables are conditioned on. The models used are typical regression models, where the specific model used depends on the measurement scale of the response variable in the regression, for example, a binary response would indicate a logistic regression model while a continuous response would indicate a linear regression model. The approach cycles through several iterations until the imputation draws appear to have converged, at which point an imputed data set is obtained. The approach is started at m random start points to obtain m multiply imputed data sets. Standard complete data methods can be applied to each imputed data set to estimate propensity scores. Combining the multiple sets of propensity scores to obtain treatment effect estimates will be discussed in Section 3.

Multiple imputation has many appealing features, notably using all available information in the data while also accounting for the uncertainty due to the presence of missing values. The MICE package in the R Statistical Software Environment is used to implement this imputation method in this paper. The MICE approach also has the advantage of producing independent imputations allowing inferences to be made with fewer imputed data sets than imputations generated through an MCMC procedure. However, there are some issues when using a standard imputation model to impute missing treatment indicators as it is the case here with a logistic regression model. The premise of analysis with propensity score matching is that we assume there are a pool of control units whose covariates lie in the same space as the treated units' covariates. In the case where we have missing treatment indicators for control units in this space, using a default logistic regression model for imputation is more likely to impute these units as treated rather than control. Not only will these lead to more inaccurate imputations but this could also lead to more imbalanced distributions between treated and matched control covariates as the pool of suitable matched controls will be reduced. This is what motivates the development of a latent class model in the next section to address this problem.

2.3 | Latent class model

The latent class model we propose develops the modeling strategy proposed by Mitra and Reiter (2011) to impute missing covariates in observational studies to obtain treatment effect estimates robust to imputation model misspecification. This is done by distinguishing control units with covariates similar to the treated units' covariates, from other control units, via a latent class indicator. The formulation thus limits the effect of outlying control units on the imputation of missing covariates and resulting treatment effect estimates. The underlying motivating principle is that an imputation model, while unsuitable over the whole covariate space, may still be a plausible model to impute missing covariates in the treated units' covariate space.

Specifically, the model assumes that the covariates distribution may be described by a two-component mixture. The first component, with $z_i = 1$, describes the distribution for those units lying in the treated units' covariate space while if $z_i = 0$ the units' covariates are assumed to follow some other distribution. The latent indicator z_i is assumed to be missing for all units where $T_i = 0$, that is, the control units, and the idea is to find those control units with covariates similar in distribution to the treated units' covariate distribution. By definition all treated units will belong to the class defined by $z_i = 1$ and so there is enough information in the data to estimate the distribution of the latent classes through MCMC.

As data sets typically contain a mix of categorical and continuous variables, Mitra and Reiter (2011) model the distribution of the data within each latent class as following a general location model (Olkin & Tate, 1961). This model decomposes the joint distribution of the variables in the data through first modeling the categorical variables jointly through a contingency table representation, and then modeling the continuous variables conditional on the categorical variables. Specifically the cell counts from a contingency table formed by cross-classifying the categorical data are described by a multinomial distribution subject to log-linear constraints (Schafer, 1997). The continuous variables are assumed to follow a multivariate normal distribution with a cell-specific mean (corresponding to which cell of the contingency table the unit lies in) and a covariance matrix pooled across the cells. Imputations of the missing covariates and the latent class indicators are drawn through a data augmentation scheme using MCMC. Details of this, and all the posterior computations, can be found in Mitra and Reiter (2011).

2.3.1 | Latent class model to impute missing treatment indicators

The modeling approach in Mitra and Reiter (2011) is solely concerned with imputing missing covariates. The treatment indicator variable is assumed to be fully observed. Here, we assume that the treatment indicator could be potentially missing and use the framework proposed in Mitra and Reiter (2011) to develop an appropriate imputation model for the missing treatment indicators.

As described above the distribution of the latent class indicators conditional on treatment can be written as,

$$p(z_i = 1|T_i = 0) = \pi^*, \tag{1}$$

$$p(z_i = 1|T_i = 1) = 1, \tag{2}$$

where π^* is some unknown probability, and each unit in the treatment group is automatically assigned to the component with $z_i = 1$. We also have the distribution of the covariates for each unit i, $\boldsymbol{x}_i$ conditional on z_i. Let us denote this by $f(\boldsymbol{x}_i|z_i)$.

We now utilize this framework to impute missing treatment indicators. In this scenario, as the treatment indicator is unknown for certain units, we must assign a probability for each unit to be allocated to treatment, and we denote this probability by π^T, so

$$p(T_i = 1) = \pi^T, \;\; i = 1, \ldots, n. \tag{3}$$

We place a beta prior distribution on π^T, so $p(\pi^T) = Be(c, d)$, where (c, d) are specified hyperparameters. Typical prior specifications for (c, d) include $c = d = 1$ (the uniform prior), and $c = d = 0.5$ (the Jeffreys prior). We use the Jeffreys prior specification here. This additional layer to the hierarchy of the model is compatible with the latent class modeling framework, and the additional unknown quantities (π^T and the missing treatment indicators) can be sampled from their full conditional distributions determined from the model. The imputation of a missing treatment indicator from its full conditional distribution is given by,

$$p(T_i = 1|z_i) = \begin{cases} \frac{\pi^T}{\pi^T+\pi^*(1-\pi^T)} & \text{for} \quad z_i = 1 \\ 0 & \text{for} \quad z_i = 0 \end{cases}. \tag{4}$$

Denote the imputed treatment indicator by $\boldsymbol{T}_{imp} = (T_{imp,1}, \ldots, T_{imp,n})'$. The conditional posterior distribution for π^T (given an imputed data set) is given by,

$$\pi^T|\boldsymbol{T}_{imp} \sim Beta\left(c + \sum_{i=1}^{n} T_{imp,i}, n + d - \sum_{i=1}^{n} T_{imp,i}\right). \tag{5}$$

One change we made to the implementation of Mitra and Reiter (2011) is to fix π^* to a value rather than specify a prior distribution for this parameter. This is due to the fact that the conditional posterior distribution of this parameter may be sensitive to the imputation of the treatment indicators and lead to very small or large values feeding back, and thus contributing to, implausible imputations of the treatment indicators. We chose a value of $\pi^* = 0.5$ as this would then have a neutral effect on the imputation of the latent class indicators, however other values, or an informative prior distribution could be considered. It is important to note that this modification does not mean imputation of the latent class indicators to group $z = 1$ happens with probability 0.5. The probability for each unit to be imputed into a latent class is still obtained from its full conditional distribution and depends on the corresponding covariate data, treatment indicator (imputed or observed), and other model parameters. The choice of 0.5 can be interpreted as giving a priori equal weight to each control unit to be in either latent class. In general, the value, or prior distribution, for π^* should be chosen to ensure plausible imputations of the missing treatment and latent class indicators. Other than π^* (which is now fixed to 0.5), the full conditional distributions to impute missing covariate values and latent class indicators, as well as updates for the rest of the model parameters remain unchanged from those derived in Mitra and Reiter (2011).

From (4) we can see if a unit is allocated to the latent class with $z = 0$, they will not be imputed to be in the treatment group. This is reasonable as the latent class $z = 0$ identifies units that are substantially dissimilar to the treated units based on their covariate values. The remaining units (with $z = 1$) are thus identified as lying within the covariate space of the treated units and thus have some probability for their treatment indicator, if missing, to be imputed to be treated.

If a unit is allocated to be in latent class $z = 1$, then the model is indicating that this unit's covariates do come from the same distribution as the treated unit's covariates, and hence there is the possibility that the unit could belong to the treatment group. The probability for that unit's treatment indicator to be imputed to treatment is determined as a function of π^T and π^*. Clearly, if $\pi^T = 0$ the probability to be imputed to treatment is 0, similarly if $\pi^T = 1$ the probability to be allocated to treatment is 1. What may not be so clear is the involvement of π^*. As π^* decreases the probability of imputing, the unit to treatment increases. This is because π^* is the probability of a control record to be imputed into latent class $z = 1$, and so given that a unit has been imputed into latent class $z = 1$, the smaller the probability of π^*, the less likely that unit is to be a control unit. As π^* increases, the probability of the unit being imputed into treatment decreases to the extent that when $\pi^* = 1$ the probability of being imputed to treatment equals π^T. This is because every control unit is now included in latent class $z = 1$, and so the latent class indicator does not give us any more information, so we can only use the information present in the model for treatment to impute missing treatment indicators.

It might appear strange that the probability of being imputed to treatment does not depend on the units' covariates $\boldsymbol{x}_i$, or their distribution $f(\boldsymbol{x}_i|z_i)$. This is because the information contained in the covariates is being implicitly included through the latent class indicator, which is imputed using the covariate data. The distribution of the latent class indicator also informs us about the distribution of the unit's unobserved covariate data. Essentially this modeling framework assumes that $x_i \perp\!\!\!\perp T_i|z_i$. This orthogonality implies that the covariate distributions, conditional on membership of a latent class, are not dependent on whether they are in the treatment or the control group. As the latent class $z = 1$ corresponds to the class comprising those units that lie in the covariate space of the treated units, this adheres to the objective of developing an imputation model that focuses its attention on the units lying within this space, thus reducing the influence of outlying control units. When the missing data mechanism only depends on $\boldsymbol{x}$, as is typically assumed, this orthogonality assumption also implies that an imputation model conditional on $\boldsymbol{x}$ and $\boldsymbol{z}$ is independent of treatment condition. Assessing properties of mechanisms that incorporate more complex dependencies, for example, on treatment, presents interesting directions for future research.

The latent class imputation model has some advantages over a standard imputation model, such as MICE, when considering the imputation of missing treatment indicators. In the latent class model, units clearly not in the covariate space of the treated units, and thus imputed to be in latent class $z = 0$, will not be imputed to be in the treatment group. In the standard imputation model every unit will have some positive probability to be imputed into treatment. In addition, the original motivation behind the development of the latent class model in Mitra and Reiter (2011) still applies: In observational studies units' covariates are spread out over a large multivariate space and so specifying plausible imputation models over this whole space is challenging. The latent class model will allow imputation of treatment indicators to be split into two groups corresponding to the latent classes, reducing the impact of outlying control units on the imputation of missing values in the covariate space of the treated units. It may then be possible to obtain more plausible imputations for the missing treatment indicators for those units in latent class $z = 1$, which may facilitate treatment effect estimates obtained from propensity score matching that are more robust to the influence of control units lying far from the treated units' covariate space.

2.4 | Estimating treatment effects with multiply imputed treatment indicators

As we are now imputing missing treatment indicators, the treatment indicators will vary from one imputed data set to the next, and so we must carefully consider strategies to estimate treatment effects in this setting. When estimating treatment effects using propensity score matching from multiply imputed data, there are two approaches to combining inferences from the multiple data sets. In the literature these have been denoted as the Within and Across methods (Hill, 2004; Mitra & Reiter, 2016). The Across method takes the propensity scores estimated from each imputed data set and averages these across the imputations before using the scores to estimate treatment effects. In contrast, the Within method uses each set of propensity scores to estimate the treatment effect within each imputed data set before averaging these. However, as we derive for the current setting, with missing treatment indicators, the Across approach is not an appropriate method to estimate treatment effects. This is because the imputed treatment indicators vary from one imputed data set to another. As a result, averaging the propensity scores across the imputations, to obtain just one set of scores, does not recognize that these scores could correspond to different treatment/control sets that vary in each imputed data set.

To construct the Within estimator, let $\boldsymbol{T}^k_{imp} = (T^k_{imp,1}, \dots, T^k_{imp,n})'$ and $\boldsymbol{X}^k_{imp}$ denote the kth imputed treatment variable and kth imputed covariate data set, respectively. In the Within method we estimate the treatment effect within each imputed data set, and then average these m treatment effect estimates. So in the kth imputed data set we estimate the propensity scores by fitting a logistic regression of $\boldsymbol{T}^k_{imp}$ on $\boldsymbol{X}^k_{imp}$ and estimate a treatment effect in the usual way (as described in Section 1); denote this by $\hat{\tau}^k$. Then the Within method estimates the treatment effect by, $\hat{\tau}^{W,m} = \frac{\sum_{k=1}^m \hat{\tau}^k}{m}$.

We now use the Within method to estimate treatment effects when we use the two modeling strategies to impute missing values and compare these estimates to the treatment effect estimate obtained from the complete case data, $\hat{\tau}^{cc}$. We first apply the methods in a simulation involving only continuous variables, before applying the methods to a real data set that includes both categorical and continuous variables.

3 | SIMULATION STUDY

We simulate a data set with 2200 units. For the first 2000 units, we simulate a covariate vector $\boldsymbol{x}_i = (x_{i1}, x_{i2})'$ so that,

$$\boldsymbol{x}_i \overset{\text{iid}}{\sim} MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, \dots, 2000$$

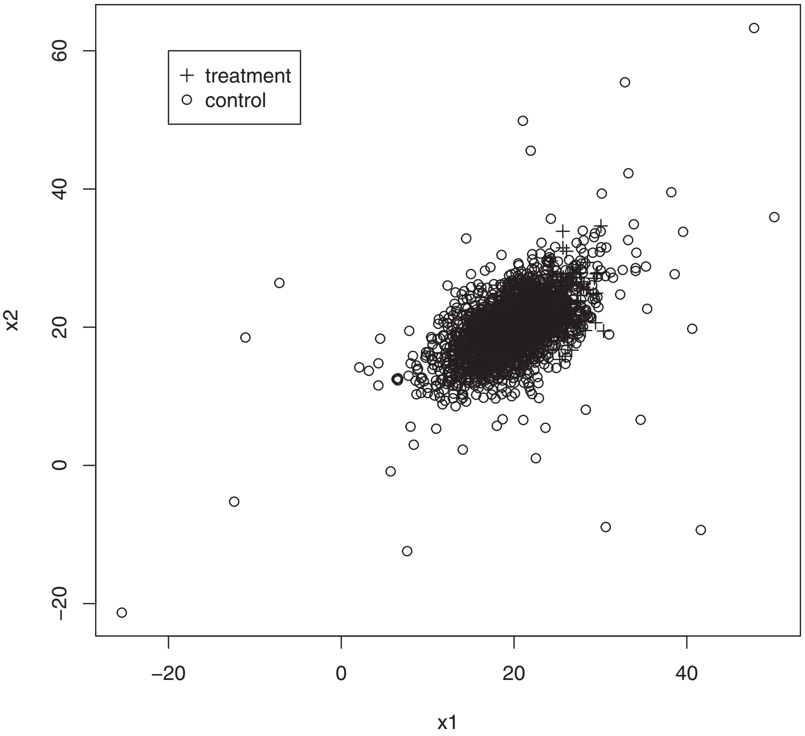

FIGURE 1 Plot of the covariate distribution in this simulation design

with $\boldsymbol{\mu} = (20, 20)'$, and $\boldsymbol{\Sigma}$ with diagonal elements 14 and off diagonal elements 7. For these 2000 units, we randomly assign 150 units to be treated from among the 300 units with the largest $\boldsymbol{x}_1$ values with the rest designated as control units. We then simulate a further 200 control units' covariate data from a multivariate t-distribution with scale matrix given by $\boldsymbol{\Sigma}$ and for units $i \in \{2001, \ldots, 2100\}$ the mean vector, $\boldsymbol{\mu} = (14, 14)'$, while for units $i \in \{2101, \ldots, 2200\}$ the mean vector, $\boldsymbol{\mu} = (26, 26)'$. We denote these groups as "outlying group 1" and "outlying group 2," respectively. This results in a data set with 150 treatment units and 2050 control units. We present a covariate plot to illustrate this in Figure 1.

We also simulate a simple response surface, y_i, for each unit i. We consider two scenarios for simulating a response surface. The first scenario (Scenario A) is where there is no treatment effect by simulating a response according to the model,

$$y_i = 2x_{i1} + x_{i2} + \epsilon_i, \text{ where } \epsilon_i \overset{\text{iid}}{\sim} N(0, 1), \quad i = 1, \ldots, 2200.$$

The second scenario (Scenario B) is where there is a treatment effect of 5 by simulating a response according to the model,

$$y_i = x_{i1} + x_{i2} + 5I(T_i = 1) + \epsilon_i, \text{ where } \epsilon_i \overset{\text{iid}}{\sim} N(0, 1), \quad i = 1, \ldots, 2200$$

and $I(\cdot)$ is the indicator function.

Now we introduce missing values in the treatment indicators. Let M_i^t be a missing data indicator for the treatment variable, so if $M_i^t = 0$ then T_i is observed, and if $M_i^t = 1$ then T_i is missing. We assign a probability for each treatment indicator in the control group to be missing conditional on its x_{i1} value using a logistic model,

$$logit(P(M_i^t = 1 | T_i = 0)) = -7 + 0.31x_{i1}.$$

This results in approximately 30% of control units with a missing treatment indicator. As the treated units lie in a comparatively smaller part of the covariate space compared to the control units, we introduce missing values in this group using an MCAR mechanism so that approximately 30% of treatment units have a missing treatment indicator. This allows both treatment and control groups to have a similar number of missing treatment indicators. Mitra and Reiter (2016) also adopted a similar approach when introducing missing values in their treatment group for similar reasons. To demonstrate that this imputation model can also simultaneously deal with missing covariate data, we also introduce missing values into

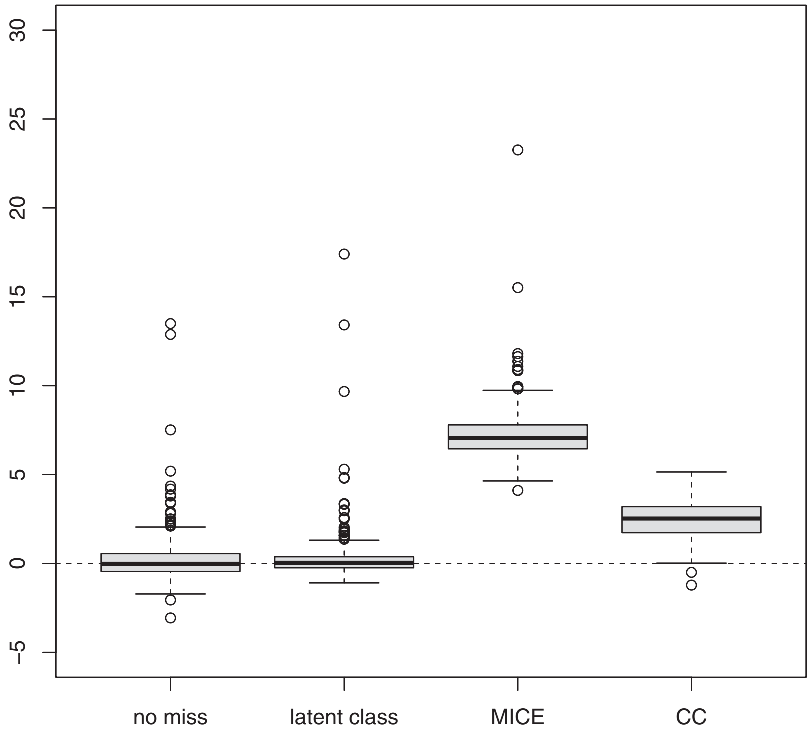

FIGURE 2 Boxplots of the estimates of the treatment effect using the latent class model to impute missing treatment indicators as well as using MICE. Also included are the treatment effect estimates without using any missing data and the treatment effect estimates from the complete case data.

units' x_{i2} using an MAR mechanism. Let M_{i2} denote the missing indicator for x_{i2} where $M_{i2} = 1$ denotes x_{i2} is missing, then missing values in x_{i2} are introduced with the mechanism,

$$logit(P(M_{i2} = 1) = -6 + 0.25x_{i1},$$

resulting in approximately 30% of units missing their x_{i2} value. The way we have simulated the data does not a priori favor either the latent class or MICE imputation model but reflects the common scenario in observational studies where units' covariates are spread out (often sparsely) over a large space.

Now we multiply impute the missing values (including the missing treatment indicators) in the data set using the latent class model described in Section 2. Note that as the covariate data are all continuous, the General location model for the covariates within each latent class reduces to a multivariate normal distribution meaning the full conditional distributions to impute missing covariates are normal distributions. We run the Gibbs sampler for 120 iterations discarding the first 20 as a burn-in, resulting in 100 imputed data sets. We noted that results stabilized quickly in the simulations, which gave confidence in making this choice. In practical situations, when the method will be applied to just one data set, the analyst could increase the number of iterations if desired. We compare results from this model with results obtained from the standard MICE imputation model using the MICE package in R. When using MICE we generate 10 imputed data sets as this is a commonly used value when using this type of approach. The MICE method can have fewer imputed data sets as it generates independent imputations as opposed to the latent class method which generates dependent imputations (from the MCMC). For each set of multiply imputed data, we estimate treatment effects using the Within method described in the previous section. We estimate propensity scores in each imputed data set with a main effects logistic regression model, and create matched control sets with nearest neighbor matching without replacement. We also estimate the treatment effect from the complete case data. We repeat this process over 500 generated incomplete data sets to compare the treatment effect estimates from the different methods.

We first consider the scenario where there is no treatment effect. We present boxplots of the treatment effect estimates from the different methods in Figure 2, the dotted line indicating the true treatment effect of 0. We see that the first boxplot, which presents estimates of the treatment effect prior to the introduction of any missing values, is close to the true treatment effect, as expected. The second boxplot presents treatment effect estimates from the latent class model and this also tends to produce estimates close to 0. The third boxplot presents the treatment effect estimates from using MICE. We see that the estimates are further from 0 compared to the estimates obtained using the latent class model. Complete case analysis here produces estimates nearer to 0 than the estimates from MICE but are still further from 0 than the latent

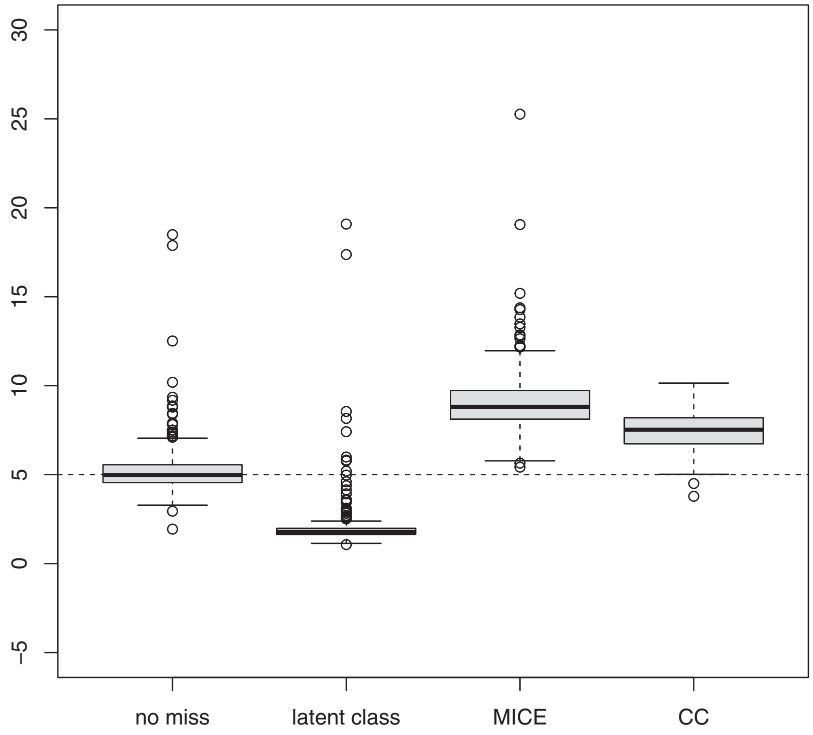

FIGURE 3 Boxplots of the estimates of the treatment effect using the latent class model to impute missing treatment indicators as well as using MICE. Also included are the treatment effect estimates without using any missing data and the treatment effect estimates from the complete case data.

class estimates. Thus, we see that by imputing missing treatment indicators using the latent class model we can reduce the bias in treatment effect estimates over other approaches.

To investigate the differences between the latent class and MICE estimates, we focus on three groups of units: the two "outlying" groups of control units, and the group of units lying in the treated units' covariate space, that is, the 300 units randomly assigned to be either treatment or control. We see on average the latent class model imputes 93.4% of units to be controls in outlying group 2 (i.e. the group comprising larger x_1 and x_2 values), while MICE imputes only 53.17%, that is, almost half are misclassified in this group by MICE. Both the latent class model and MICE impute few units to be treated in outlying group 1 (i.e., the group comprising small x_1 and x_2 values), with 98.8% and 95.7% of units imputed to be controls, respectively. We also see that MICE imputes many more controls lying in the treated units' covariate space to be treated compared to the latent class model, with misclassification rates among missing values in this matched control group of 58.86% (MICE) versus 9.61% (latent class), respectively. When it comes to the imputation of treated units with missing treatment indicators, on average fewer missing values are imputed to be treated using the latent class model than with MICE, with on average approximately 41 units imputed to be control with the latent class model versus 20 units with MICE. However, as the number of missing treatment indicators is relatively small (46 on average approximately) compared to the number of missing controls in the same region (108 on average approximately), coupled with the true treatment effect being zero, this does not significantly effect treatment effect estimates obtained using the latent class model. In contrast, using MICE results in a substantial inflation in the number of treatment indicators, both in the region of the covariate space where the treated units lie, as well as in outlying group 2. This thus reduces the pool of appropriate matched controls to be used in the comparison group to infer treatment effects, and results in matched controls being selected with covariates more dissimilar to the treated units' covariates. The resulting covariate imbalance introduced through using MICE results in the biases in the treatment effect seen in Figure 2.

We also present results from Scenario B, where there is a treatment effect of 5. As before we present boxplots of the treatment effect estimates in Figure 3; the dotted line now indicates the true treatment effect of 5. The estimates from the fully observed data are again close to the true treatment effect. We now see that when we impute the missing data using the latent class model we obtain treatment effect estimates that are smaller than 5 in general, with a bias of −3. However, these estimates are closer to 5 than the estimates obtained using MICE, which has a bias of 4. The latent class approach also has a lower MSE compared to MICE (10.6 versus 18.7, respectively). There are thus still some gains, albeit modest, when using the latent class model over MICE in Scenario B, although the gains are more evident in Scenario A. The complete case estimates have a similar distance to the true treatment effect as the latent class estimates.

We perform similar investigations to explain the results observed. We see that, as in Scenario A, MICE over imputes missing treatment indicators, that is, many missing controls are imputed to be treatment, with on average 34.5% of controls in outlying group 2 and 26.1% of controls lying in the covariate space of the treated units with missing treatment indicator imputed to be treatment. However, these numbers are smaller in Scenario B due to the fact we condition on the response in the imputation models, and the response here imparts important information about treatment status due to the presence of a treatment effect in this scenario. The latent class model performs similarly to Scenario A with fewer controls in outlying group 2 (6.24%) and in the covariate space of the treated units (8.33%) imputed to be treated as compared to MICE. As with Scenario A, both MICE (3%) and the latent class method (1.1%) impute few controls in outlying group 1 to be treated. However, as with Scenario A, there are few treated units with missing treatment indicators imputed to treatment (5.17 out of 45.5 units on average) with the latent class model, while MICE performs comparatively better (31.44 out of 45.5 units on average). These misclassified imputed treatment indicators (classified now as controls) are likely to be selected to the matched control group and used for comparisons with the treatment group, this explains the negative bias of −3 we see in the treatment effect estimate using the latent class model. The relatively small number of treated subjects with missing treatment status (45.5 on average out of 150 treated units) means the bias using the latent class approach is still smaller compared to the bias of 4 when using MICE, although the difference in performance is smaller in this scenario.

When considering the complete case estimates, we see the estimates are simply shifted by exactly 5 units between both scenarios, as expected, with the only difference between the two scenarios being the omission/introduction of a treatment effect, respectively. As a result, in both scenarios, the bias when using complete cases is 2.46, and the variance is 1.15. The reason for the bias is due to a lack of suitable control units representative of the treated group through the omission of units with missing treatment indicators, many of which are controls that would ordinarily have been selected to the matched control group had they been observed. The larger interquartile range of the complete case estimates compared to the imputation methods, as seen by the boxplots in Figures 2 and 3, is due to the smaller sample size resulting from a complete case analysis, a common phenomenon and drawback associated with this approach.

The results presented above, and subsequent inspection of these, illustrate the potential benefits of imputing missing treatment indicators using the latent class model over other approaches. However, different or more complex settings may yield changes to the relative performances seen here, and this is an interesting topic for future investigation.

4 | APPLICATION TO A STUDY ON DETERMINANTS OF CHILDREN'S COGNITIVE DEVELOPMENT

We now apply the latent class model to impute missing values and perform propensity score matching in a study that investigated effects of various factors on a children's cognitive development. The data are a subset of the 1979 National Longitudinal Survey of Youth, commonly referred to as the NLSY79. This data set was also analyzed by Mitra and Reiter (2011) to multiply impute missing covariate values, but did not consider the additional complication of missing treatment indicators. We refer readers to Mitra and Reiter (2011) for more details concerning the data set. This data set is used only to illustrate the performance of the imputation models described and not to draw any definitive conclusions concerning causal effects; the common complications of not missing at random mechanisms and unmeasured confounding could well apply here.

4.1 | Description of variables

We consider 15 variables in total. These include: the child's race (Hispanic, black or other), child's sex, breastfeeding indicator (0 if less than 24 weeks, 1 otherwise), two variables indicating whether the spouse or grandparents were present at birth, an indicator for if the child was born premature (0 if 0 weeks premature, 1 for > 0 weeks), the number of weeks that the mother worked in the year prior to giving birth (not worked at all, worked between 1 and 47 weeks, worked 48–51 weeks, and worked all 52 weeks), the number of years between 1979 and when the mother gave birth, mother's intelligence as measured by an armed forces qualification test, mother's highest educational attainment, child's birth weight, the number of days that the child spent in hospital, the number of days that the mother spent in hospital, family income, and a measure of child's cognitive development through the Peabody individual assessment test math (PIATM) score administered to children at 5 or 6 years of age. We did not consider mother's race due to its likely similarity with child's race. Following Mitra and Reiter (2011), we applied Box–Cox transformations to several continuous variables to facilitate imputation modeling. We refer readers to Mitra and Reiter (2011) for more details on these.

The response variable, y, is the PIATM score. However, the treatment variable considered in Mitra and Reiter (2011), breastfeeding, had only 5.6% of its values missing. As a result we instead consider family income as our treatment variable, which has 22.3% missing values. Specifically, we define a unit to be treated if it lies in the top decile of the family income variable. There is evidence to suggest that family income plays a role in children's cognitive development (Cooper & Stewart, 2021) and a naive difference in means of PIATM scores in the treatment and control groups gives a difference of approximately 8.38 points.

We include only first-born children in the analysis to avoid complications due to birth order and family nesting. We also discard 4977 units with a missing PIATM; this is reasonable under missing at random assumptions, which may not be true in practice. We do not consider methods for handling the missing outcome data in the analysis here, as the cases with complete outcome data suffice for our purposes: to examine the impact on treatment effect estimates after using the latent and MICE imputation models, respectively, as well as comparing these to a complete case analysis. The resulting data comprise 2531 youths, of whom 197 are treated. Three covariates are completely observed in the study, and eight covariates have missing data rates of less than 10%. The two covariates with the largest rates of missing data are mother's highest education attainment (23.7%) and the number of weeks that the mother worked in the year prior to giving birth (22.3%). There are 991 units with complete data on all covariates, of whom 72 are treated.

4.2 | Simulation based on the complete cases

Before we consider analyzing the full data set, we evaluate the performance of the imputation models in a simulation involving the 991 complete cases. As this is a much smaller sample size, we first increase the sample size to the original data sample size through univariate resampling of the rows of the complete case data. This results in a fully observed data set of 2531 units. We then introduce missing data back into the fully observed data to reflect the distribution of missing data patterns present in the original sample. We then run the latent class model and MICE to impute missing values and compare the treatment effect estimates to the treatment effect estimate from the original fully observed data. To obtain a treatment effect from the fully observed sample, we estimate propensity scores using a main effects logistic regression model on all the covariates, and create a matched control set using nearest neighbor matching without replacement, which yields a treatment effect estimate of approximately 2.68. We repeat this process 100 times to generate 100 incomplete data sets from the fully observed sample.

We multiply impute the missing values using both the latent class model as well as using MICE. For the latent class model, following Mitra and Reiter (2011), we use a General Location latent class model with a main effects only log linear model for the categorical variables, making use of the convenient mix package in R to greatly facilitate computations. We use the transformed continuous variables that were suggested by the Box–Cox procedure, and relate the within-category means using a linear model with main effects of the categorical variables. We run the Gibbs samplers for 220 iterations after discarding an initial 20 as burn-in to create 200 multiply imputed data sets. We note the analyst could run the model for more iterations if this was desired. For the MICE method we again generate 10 imputed data sets. We estimate treatment effects from the multiply imputed data using the Within method, estimating the propensity scores in each imputed data set with a main effects logistic regression, and creating matched control sets by nearest neighbor matching without replacement.

Figure 4 presents boxplots of the treatment effect estimates arising from the latent class and MICE imputation models across the 100 replications. The horizontal dashed line indicates the treatment effect from the fully observed data. The treatment effect estimates from the latent class model tend to be slightly closer to the treatment effect estimate from the fully observed data than the corresponding treatment effect estimates obtained using MICE. What is striking is the much greater variability in the treatment effect estimates from the complete case analysis, which shows the potential drawbacks of this approach.

Based on these simulations we see there are only some small potential benefits in using the latent class model for imputation over using MICE. It is however interesting to see the potential drawbacks a complete case analysis could have here, which is a common way to handle missing treatment indicators. As the process of introducing missing values back into the complete data will naturally be MCAR, the estimates from a complete case analysis would typically expect to be unbiased but with the potential to have a high variance due to the loss of data. Indeed we see evidence for this in Figure 4 with the boxplot from the complete case approach having the greatest spread. It is worth also noting though that complete

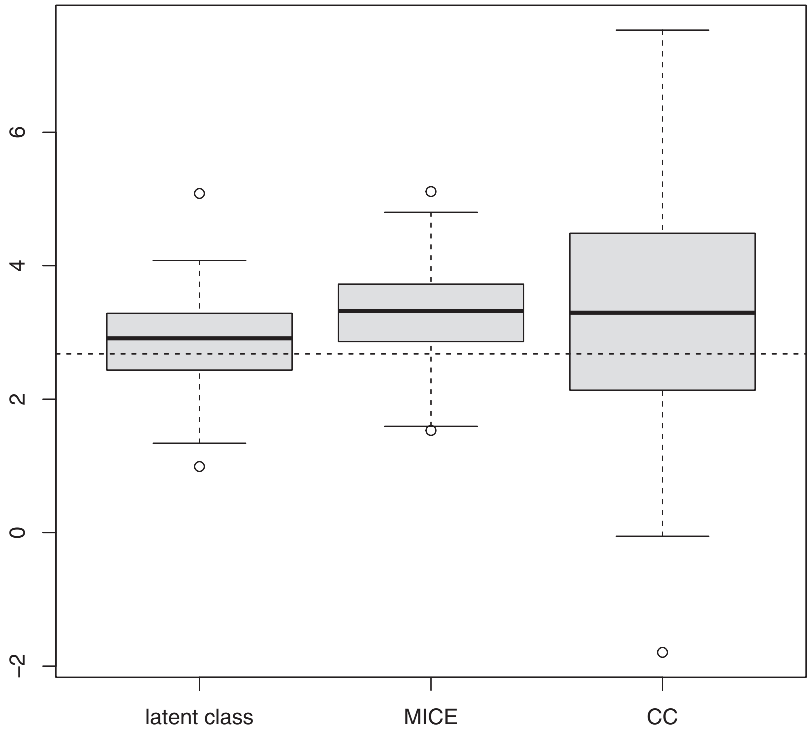

FIGURE 4 Boxplots of the estimates of the treatment effect using the latent class model to impute missing treatment indicators as well as using MICE. Also included are the treatment effect estimates from a complete case analysis and the treatment effect from the fully observed data (horizontal dashed line).

case analysis could still result in biases if the complete cases do not allow a sufficiently representative control group to be drawn upon when making inferences about the treatment effect, and we see some evidence for this occurring in Figure 4 as well.

We conjecture that the missing data mechanism being generated in a simpler way, as compared to the simulations in Section 3, is a key factor behind why both imputation methods perform similarly well, and inline with what we expect to see. The simpler missing data mechanism means it is easier for the imputation models to plausibly impute the missing treatment indicators and thus ensure treatment effect estimates are not adversely affected by their presence. It may be the case that a more complex missing data generation process would lead to greater differences in performance, for example, as seen in Simulation Scenario B. One way to do this would be to consider introducing differential rates of missingness in the treated and control units' treatment indicators, using different missing data mechanisms for each. However, this would risk becoming very subjective. Our goal in this section is to design a simulation experiment that best reflects the real data setting, while also not a priori favoring a particular method to deal with the missing values. Hence, we feel that the simulation design used best adheres to these objectives.

For some data sets, when implementing the latent class model, we occasionally encountered an issue with a latent class allocation that resulted in too few units allocated to a class to estimate all the model parameters. This occurred when there was an insufficient number of units to encompass the levels of each categorical variable. In these situations, we randomly switched units from the other latent class, with the necessary covariate attributes, to the affected latent class. The number switched was small, and determined by a specified minimum threshold for the number of observations within each level of a categorical variable that must be in each class. In the results reported, the threshold was 4. Where issues related to underrepresentation still occurred, we simply discarded that replication and generated another data set (this could also have been another way to address this problem on its own). While the ultimate effect on the results is small, we note this is an interesting finding that could be an issue in general when applying latent class models to data sets with lots of categorical variables and levels. Some other strategies that could be explored are specifying informative priors for all model parameters, or restricting draws of latent class indicators to satisfy thresholding properties. The former approach can unfortunately not be implemented within the mix package in R that facilitated implementation of the latent class model. The latter approach corresponds to imposing a data-dependent prior that has been suggested in the literature (Diebolt & Robert, 1994; Wasserman, 2000). We tried implementing this approach but found this became computationally infeasible in some cases. Addressing this problem in general presents an interesting topic for future research.

4.3 | Application to the full data

We now apply the latent class and MICE approaches to impute missing values in the full data set using the same settings as above. From the latent class model we obtained a treatment effect estimate of 2.72 while MICE gave a treatment effect estimate of 3.71. We thus see that there is only a very small difference in the estimates in this instance. These differences are not of great practical significance, and partly this could be attributed to the fact that the need for a latent class is not exceptionally strong. Nevertheless, we see that the implementation of the latent class model does not appear to introduce additional bias into the estimates of the treatment effect, at least for the simulations considered here. It is also worth noting that the treatment effect estimate from a complete case analysis here is 3.58, which is similar to the estimates obtained from both imputation methods.

5 | CONCLUDING REMARKS

In this paper, we consider multiply imputing missing treatment indicators when estimating treatment effects in observational studies through propensity score matching. We propose a novel latent class model to multiply impute missing treatment indicators. The latent class model makes use of the fact that the distribution of the treated units' covariates are different than the majority of the control units' covariate distribution, and this can help inform us about which control units with missing treatment indicators are more likely to be in the treatment group than the others. Through simulations we see that there are potential gains when we use the latent class model for imputation in reducing bias in treatment effect estimates compared to using the MICE approach for imputing missing treatment indicators. An important finding in this paper is the potential cost of simply ignoring this issue through performing a complete case analysis. Through the simulations in this paper we can see that a complete case analysis results in estimates of the treatment effect that have a greater variability compared to estimates from the imputed data. A complete case analysis could also increase bias in the treatment effect estimates.

In all analyses reported in this paper, we included the response variable in our imputation models. This is to help specify the most appropriate imputation model for the missing values, with the response providing valuable information for the imputations. However, we recognize alternative views that suggest the response should not be included in the imputation model (D'Agostino Jr. & Rubin, 2000). It would be interesting to explore this issue further, that is, consider if there were any scenarios when including the response would be clearly inappropriate.

It would be interesting to consider whether the method proposed here could be adapted to deal with more general treatment conditions. Propensity score methods have been developed for categorical and ordinal treatment conditions. Imai and Van Dyk (2004) review these and generalize this further to include, among other types, continuous treatments. In principle, the methodology developed here could be extended to deal with multiple treatments although there are some important computational and theoretical considerations. These include determining the number of latent classes, for example, whether one extra class per additional treatment group is required, as well as considering how the imputation model could be best constructed to accommodate the multiple possible treatment values.

In general, even in the binary treatment setting, it would also be interesting to explore whether there would be any benefits to extending this model to incorporate several latent classes. This could involve splitting the latent class $z = 0$ into several latent classes, so that units could belong to the latent class corresponding to the treated units covariate distribution, or some other class corresponding to another distribution of covariates. Thus, we would be recognizing the possibility of there being several groups of control units, each with a different covariate distribution, rather than there only being two groups in the data. Similarly, we could also consider splitting the latent class $z = 1$ into several latent classes recognizing that there may be several groups of units, each lying in the treated units' covariate space, with different covariate distributions. The novel twist of this approach, that of considering certain control units to belong to the covariate distribution of the treated units, remains the same. Now controls might belong to one of several covariate distributions corresponding to a particular group of treated units. A natural question then arises of how many latent classes would be required. We could conduct a sensitivity analysis, but an interesting approach could be to specify the number of latent classes as unknown through a Dirichlet process distribution. This would allow us to form covariate clusters upon which the imputation models would be based.

The positivity assumption also merits careful consideration in this setting. Positivity is mostly a concern with categorical covariates, and requires every exposure to have a positive probability for all strata/levels of each covariate. In our setting, as we only have two exposure groups (treatment and control), positivity requires positive probability for a unit to be in treatment/control for each level of a categorical covariate. In practice, positivity will be satisfied provided there are a mix of treated and control units within each level of a categorical variable. While the latent class method does form a class comprising only control units, this does not violate positivity as this is only for imputation of missing values and when estimating treatment effects the whole sample is used. However, the problem could occur if imputation of treatment indicators results in only imputed treated or control units for a level of a categorical variable. This can be avoided if a restriction is imposed in the imputation model to ensure the imputed treatment indicator comprises at least one treated and one control unit in each level of a categorical variable. Investigating this further is an interesting area of future research.

It is clear that more attention needs to be given to this area. In particular, considering the situations where imputation of missing treatment indicators would be most beneficial and whether there were any situations where imputation could have significant drawbacks. Further it would be interesting to investigate whether diagnostic tools could be developed to determine whether missing treatment indicators are likely to affect inference of the treatment effect. These are all interesting areas of future research.

ACKNOWLEDGMENTS

We would like to thank Professor Jennifer Hill for providing the data used in Section 4, and Professor Jerome Reiter for his helpful comments and suggestions in the preparation of this paper. We are also very grateful to the associate editor and a reviewer for very helpful and constructive suggestions.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

The study data considered in the manuscript is secondary data, taken from the National Longitudinal Survey of Youth, and can be accessed as a public use file from the U.S. Bureau of Labor Statistics website.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Robin Mitra https://orcid.org/0000-0001-9584-8044

REFERENCES

Ahn, J., Mukherjee, B., Gruber, S. B., & Sinha, S. (2011). Missing exposure data in stereotype regression model: Application to matched case–control study with disease subclassification. *Biometrics*, *67*(2), 546–558.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment group in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107.

Breger, T. L., Edwards, J. K., Cole, S. R., Westreich, D., Pence, B. W., & Adimora, A. A. (2020). Two-stage g-computation: Evaluating treatment and intervention impacts in observational cohorts when exposure information is partly missing. *Epidemiology*, *31*(5), 695–703.

Cho, Y. B., Lee, K., Suh, K., Kim, Y., Yoon, J., Lee, H., S., H., & Park, B. (2007). Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Journal of Gastroenterology & Hepatology*, *22*(10), 1643–1649.

Cooper, K., & Stewart, K. (2021). Does household income affect children's outcomes? A systematic review of the evidence. *Child Indicators Research*, *14*(3), 981–1005.

da Veiga, P. V., & Wilder, R. P. (2008). Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Maternal and Child Health Journal*, *12*(2), 194–203.

D'Agostino Jr, R. B., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, *95*(451), 749–759.

Demissie, S., LaValley, M. P., Horton, N. J., Glynn, R. J., & Cupples, L. A. (2003). Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Statistics in Medicine*, *22*(4), 545–557.

Diebolt, J., & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(2), 363–375.

Hanley, J. A., & Dendukuri, N. (2009). Efficient sampling approaches to address confounding in database studies. *Statistical Methods in Medical Research*, *18*(1), 81–105.

Hill, J. (2004). *Reducing bias in treatment effect estimation in observational studies suffering from missing data*. Working paper 04-01, Columbia University Institute for Social and Economic Research and Policy (ISERP).

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In A. Gelman, & X. L. Meng (Eds.) *Applied Bayesian modeling and causal inference from an incomplete-data perspective*. Wiley.

Hullsiek, K. H., & Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics (Oxford)*, *3*(2), 179–193.

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854–866.

Kennedy, E. H. (2020). Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics*, *16*(1).

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J., Bell, M. L., Goetghebeur, E., Hogan, J. W., & Carpenter, J. R. (2021). Framework for the treatment and reporting of missing data in observational studies: The treatment and reporting of missing data in observational studies framework. *Journal of Clinical Epidemiology*, *134*, 79–88.

Leon, A. C., Demirtas, H., Li, C., & Hedeker, D. (2012). Two propensity score-based strategies for a three-decade observational study: Investigating psychotropic medications and suicide risk. *Statistics in Medicine*, *31*(27), 3255–3260.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*(19), 2937–2960.

Mebane Jr, W. R., & Poast, P. (2013). Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Political Analysis*, *21*(2), 233–251.

Mitra, R., & Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*, *30*(6), 627–641.

Mitra, R., & Reiter, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*, *25*(1), 188–204.

Molinari, F. (2010). Missing treatments. *Journal of Business and Economics Statistics*, *28*(1), 82–95.

Nguyen, V. C. (2012). Program impact evaluation using a matching method with panel data. *Statistics in Medicine*, *31*(6), 577–588.

Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, *32*(2), 448–465.

Park, G. S., Wong, W. K., Oh, M., Khanna, D., Gold, R. H., Sharp, J. T., & Paulus, H. E. (2007). Classifying radiographic progression status in early rheumatoid arthritis patients using propensity scores to adjust for baseline differences. *Statistical Methods in Medical Research*, *16*(1), 13–29.

Qu, Y., & Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine*, *28*(9), 1402–1414.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B-Methodological*, *53*, 597–610.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.

Sartori, N., Salvan, A., & Thomaseth, K. (2005). Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. *Computational Statistics & Data Analysis*, *49*(3), 937–953.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, *17*, 546–555.

Shardell, M., & Hicks, G. E. (2014). Statistical analysis with missing exposure data measured by proxy respondents: A misclassification problem within a missing-data problem. *Statistics in Medicine*, *33*(25), 4437–4452.

Shortreed, S. M., & Forbes, A. B. (2010). Missing data in the exposure of interest and marginal structural models: A simulation study based on the Framingham Heart Study. *Statistics in Medicine*, *29*(4), 431–443.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21.

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.

Wasserman, L. (2000). Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(1), 159–180.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*, 826–833.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

Williamson, E., Forbes, A., & Wolfe, R. (2012a). Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Statistics in Medicine*, *31*(30), 4382–4400.

Williamson, E., Morley, R., Lucas, A., & Carpenter, J. (2012b). Propensity scores: From naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, *21*(3), 273–293.

Woo, M.-J., Reiter, J. P., & Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine*, *27*(19), 3805–3816.

Zhang, Z., Liu, W., Zhang, B., Tang, L., & Zhang, J. (2016). Causal inference with missing exposure information: Methods and applications to an obstetric study. *Statistical Methods in Medical Research*, *25*(5), 2053–2066.

Zhu, J., & Raghunathan, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, *110*(511), 1112–1124.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mitra, R. (2022). A latent class model to multiply impute missing treatment indicators in observational studies when inferences of the treatment effect are made using propensity score matching. *Biometrical Journal*, 1–17. https://doi.org/10.1002/bimj.202100284