Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2024;16(3):129-39

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA DOI: 10.5336/biostatic.2024-105388

A Methodological Study on Estimating Propensity Scores with Missing Semi-Continuous Covariate Data: Application to Maternal Drinking and Childhood Cognition

Eksik Kısmi-Sürekli Kovaryat Verileri ile Eğilim Puanlarının Tahminine Yönelik Yöntemsel Bir Çalışma: Anne İçkisi ve Çocukluk Bilişi Üzerine Uygulama

Tuğba AKKAYA HOCAGİL[a], Richard J. COOK[b], Sandra W. JACOBSON[c], Joseph JACOBSON[c], Louise M. RYAN[d,e]

[a]Ankara University Faculty of Medicine, Department of Biostatistics, Ankara, Türkiye
[b]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada
[c]Department of Psychiatry and Behavioral Neurosciences, Wayne State University School of Medicine, Detroit, USA
[d]School of Mathematical and Physical Sciences, University of Technology Sydney, Ultimo, NSW, Australia
[e]Harvard T.H. Chan School of Public Health, Boston, MA, USA

ABSTRACT Objective: This study adapts propensity score methodology for estimating causal effects in observational studies, particularly addressing the challenges of missing semi-continuous covariate data in environmental epidemiology. **Material and Methods:** Data were drawn from the Detroit Longitudinal Cohort Study, which examines prenatal alcohol exposure's impact on child cognitive development. The dataset includes maternal self-reports of alcohol and drug use during pregnancy, alongside biological assay results. A significant portion of the covariates, such as maternal substance use, exhibit a semi-continuous distribution with excess zero values and a long tail. Missing data in these covariates pose a risk to valid causal inference. To address this, we used the R package MICE for multiple imputation, incorporating maternal characteristics, socioeconomic indicators, and child neurodevelopmental outcomes. Additionally, a two-part modeling approach accounted for the distinct zero-inflated nature of the covariates. Misclassification correction techniques reconciled discrepancies between biological assays and maternal self-reports, particularly for illicit drug use, by adjusting sensitivity and specificity during the imputation process. Propensity scores for gestational alcohol exposure were estimated using the imputed datasets to ensure balanced covariates across exposure groups. **Results:** Our method performed well, particularly in scenarios with high percentages of zeros and missing observations in the semi-continuous covariates. **Conclusion:** This approach provides robust estimates of propensity scores, enhancing causal inference in studies involving maternal behaviors and childhood cognition.

Keywords: Multiple imputation for semi-continuous covariates; propensity score with partially observed covariates; Detroit Longitudinal Cohort Study; two-part structure; misclassification

ÖZET Amaç: Bu çalışma, gözlemsel çalışmalarda nedensel etkilerin tahmin edilmesinde eğilim skor metodolojisini uyarlamaktadır, özellikle çevresel epidemiyolojide kısmi-sürekli kovaryat verilerinin eksikliği ile ilgili zorluklara odaklanmaktadır. **Gereç ve Yöntemler:** Veriler, gebelikte alkol maruziyetinin çocuk bilişsel gelişimi üzerindeki etkisini inceleyen Detroit Boylamsal Kohort Çalışmasından alınmıştır. Veri seti, gebelik sırasında anne tarafından bildirilen alkol ve ilaç kullanımı ile biyolojik test sonuçlarını içermektedir. Kovaryatların önemli bir kısmı, anne madde kullanımı gibi fazla sıfır değeri ve uzun kuyruklu dağılımlar sergileyen kısmi-sürekli bir dağılıma sahiptir. Bu kovaryatlarda eksik veriler, geçerli nedensel çıkarım için risk oluşturur. Bu durumu ele almak için anne özellikleri, sosyoekonomik göstergeler ve çocuk nörogelişimsel sonuçları içeren çoklu imputasyon için R paketi MICE kullanılmıştır. Ayrıca, kovaryatların sıfır şişirilmiş doğasını hesaba katmak için iki parçalı modelleme yaklaşımı uygulanmıştır. Yanıltıcı sınıflandırma düzeltme teknikleri, biyolojik testler ile anne raporları arasındaki uyumsuzlukları, özellikle yasa dışı ilaç kullanımı için, imputasyon sürecinde duyarlılık ve özgüllük ayarlamaları yaparak uyumlu hâle getirmiştir. Gestasyonel alkol maruziyeti için eğilim skoru, impute edilmiş veri setleri kullanılarak tahmin edilmiştir ve bu sayede maruziyet grupları arasında dengeli kovaryatlar sağlanmıştır. **Bulgular:** Yöntemimiz, özellikle kısmi-sürekli kovaryatlarda yüksek sıfır oranları ve eksik gözlemler bulunan senaryolarda iyi performans göstermiştir. **Sonuç:** Bu yaklaşım, maternal davranışlar ve çocuk bilişi ile ilgili çalışmalarda nedensel çıkarımı geliştiren sağlam eğilim skoru tahminleri sunmaktadır.

Anahtar kelimeler: Kısmi-sürekli kovaryatlar için çoklu imputasyon; kısmi gözlemlenen kovaryatlar ile eğilim skoru; Detroit Boylamsal Kohort Çalışması; iki parçalı yapı; yanıltıcı sınıflandırma

TO CITE THIS ARTICLE:
Akkaya Hocagil T, Cook RJ, Jacobson SW, Jacobson J, Ryan LM. A methodological study on estimating propensity scores with missing semi-continuous covariate data: Application to maternal drinking and childhood cognition. Turkiye Klinikleri J Biostat. 2024;16(3):129-39.

Correspondence: Tuğba AKKAYA HOCAGİL
Ankara University Faculty of Medicine, Department of Biostatistics, Ankara, Türkiye
E-mail: akkaya.tugba@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 06 Sep 2024 ***Received in revised form:*** 01 Jan 2025 ***Accepted:*** 07 Jan 2025 ***Available online:*** 14 Jan 2025

2146-8877 / Copyright © 2024 by Türkiye Klinikleri. This in an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Epidemiological studies on exposures like alcohol, smoking, or environmental factors must control for confounders to make valid health inferences. Propensity score methods, a flexible alternative to regression analysis, estimate the probability of receiving an exposure based on observed covariates, balancing treated and untreated groups. These methods can be applied to binary, continuous, or semi-continuous exposures, using techniques like matching, stratification, and inverse probability weighting for binary exposures, and regression adjustment or inverse probability weighting for continuous exposures.[1,2] A key assumption in propensity score analyses is the absence of unmeasured confounding, meaning all potential confounders must be included in the model. To meet this assumption, Rubin and Thomas (1996) recommend including a wide range of confounders in the propensity score model.[3] However, missing data in the covariates poses a challenge. Complete case analysis can reduce sample size, lowering statistical power, increasing standard errors, and introducing bias if the data are not missing completely at random (MCAR). A more sophisticated alternative is multiple imputation (MI), which effectively handles missing data in various contexts.[4-6]

Several studies have explored using multiple imputation (MI) to create complete datasets for estimating propensity scores.[7,8] Donna et al. (2000) compared MI with generalized boosted modeling and found that MI, particularly with a missingness pattern, outperformed other methods by accounting for the added uncertainty in imputing missing data.[9] While many statistical packages for MI exist, they typically rely on standard regression models like logistic or linear regression, which may not suit covariates with different distributions. This paper addresses challenges in environmental and behavioral epidemiology, where variables are often long tailed or semi-continuous. For example, zero values can indicate abstinence (e.g., no exposure), which differs qualitatively from non-zero values representing varying exposure levels. Abstinence may be linked to distinct socio-environmental and personal confounders, and such long-tailed, semi-continuous variables are common when measuring environmental contaminants.

Several approaches have been recommended for handling incomplete semi-continuous data. One useful method is multiple imputation with predictive mean matching (PMM), which preserves the structure of the variable by sampling from similar predictive mean values.[10] Because it preserves the structure of the variable by sampling from observations with similar predictive mean values. Another approach is log-transforming the skewed variable before imputation and back-transforming it to its original scale.[11-13] However, this treats semi-continuous variables as continuous, which can be problematic. A better method imputes missing data in two steps: first, imputing the probability of exposure, then imputing the amount of exposure for those predicted to be exposed. This approach has been applied to cost data, though the initial binary step treated the data as normally distributed, rounding imputed values to 0 or 1, using a cutoff value of 0.5.[14] Su et al. developed an algorithm for semi-continuous data imputation, implemented in the R package “mi”.[15] In this paper, we apply a similar algorithm using standard MI software (e.g., MICE), with additional steps to handle misclassified covariates. We propose a multiple imputation scheme that effectively addresses skewed and semi-continuous variables while incorporating supplementary information from biological assays, improving the plausibility of imputations—something PMM does not offer.

The paper is structured as follows: The next subsection introduces our motivating example. Section 2 explains how the R package MICE can be adapted to impute missing semi-continuous data using a two-part model and handle misclassification, also demonstrating the estimation of a propensity score for a continuous exposure. Section 3 presents results from a simulation study comparing our imputation scheme with predictive mean matching and a normal imputation routine. Section 4 discusses findings from the Detroit Longitudinal Cohort Study. Section 5 reviews the strengths and limitations of our approach and suggests future directions.

MATERIAL AND METHODS

MOTIVATING EXAMPLE

Prenatal alcohol exposure (PAE) is linked to cognitive and behavioral deficits, but the dose-response relationship remains unclear. While pregnant women are advised to abstain from alcohol, the effects of low-

level exposures are still not well understood, complicating diagnosis and treatment of affected children. This issue is further complicated by studies often relying on clinic-referred children, where accurate data on maternal alcohol use is lacking.

The Detroit Longitudinal Cohort Study follows children with varying levels of PAE, tracking participants from birth to age 19. The study began with 480 pregnant African American women and collected alcohol use data through timeline follow-back interviews, along with information on other substance use, smoking, demographics, maternal health, and home environment. Neuropsychological tests, including IQ, memory, executive function, and academic achievement, were administered from infancy to adulthood.

Ethics approval for this study was granted by the University of Waterloo Research Ethics Office (approval number 40535, dated December 19, 2018), ensuring compliance with ethical standards and the principles of the Helsinki Declaration.

PROPENSITY SCORE ESTIMATION IN THE PRESENCE OF MISSING COVARIATES

Suppose we are interested to relate a continuous exposure variable, T, to an outcome Y, after adjusting for a set of potential confounders, X_1, X_p. Our analysis will be based on the use of propensity score methods, as they have been adapted to the context of continuous exposure variables.[1] We assume that there are no missing values for the Y and T, but in the context where some of the Xs are missing, analysis needs to follow the following steps:

i. Apply MICE to impute the desired number of completed datasets. In practice, it is often recommended to generate 10 imputed datasets.

ii. For each completed dataset, do the following:

- Fit a model that predicts T as a function of all available potential confounders, including interactions and non-linear terms, as needed. For each subject, obtain their predicted value $\hat{T}$
- Fit the outcome model: $Y^{16} = \omega_0 + \omega_1 T + \omega_2 \hat{T} + \varepsilon$.

iii. After repeating step 1-2 for each imputed data set, we use Rubin's rules to combine the estimated ω_1 and their standard errors obtained from each imputed dataset.[16]

In principle, these steps are straightforward. However, a concern arises when one or more of the Xs have a skewed, semi-continuous structure. In our motivating study, this situation occurs with variables related to prenatal use of cocaine, marijuana, opiates, and tobacco use during pregnancy. In the next section, we will discuss how to adapt the R package MICE to impute semi-continuous covariates while accommodating misclassification in the imputation process.[17]

MULTIPLE IMPUTATION BY CHAINED EQUATIONS FOR SEMI-CONTINUOUS VARIABLES

To simplify the discussion, we drop the subscript that distinguishes different confounders and consider a variable X with a two-part structure, where X=0 denotes zero exposure, and nonzero values represent the amount of exposure. To handle missing values in X, we can think of X as the product of two ancillary variables, W a binary indicator of exposure, and Z, an ancillary continuous variable. This approach is described and implemented in the R package mi.[15] While it may seem unnecessary to interpret Z when W=0, it is not crucial since our primary interest is in the product X=Z*W. The elegance of this framework lies in its adaptability within MICE, as it allows for the imputation of missing values for the log-transformed, allowing for the imputation of missing values for log-transformed Z and W. During the analysis phase, only the expression X=exp(Z)*W is utilized. The specific pattern of missingness for W and Z is outlined as follows:

W = 1, if X > 0
= 0, if X = 0
= missing, if X is missing,

Z = X, if X > 0
= missing, if X = 0
= missing, if X is missing.

CORRECTING FOR MISCLASSIFICATION IN THE MULTIPLE IMPUTATION BY CHAINED EQUATIONS

Misclassification can occur in epidemiological studies when participants have a reason to provide false responses, particularly with sensitive questions, such as those related to illicit drug use. If not corrected, this misclassification can lead to biased results and incorrect conclusions.

In our motivating study, routine urine toxicology screen results revealed that some subjects who reported no illicit drug use at their prenatal visits tested positive for illicit drugs. We used this supplementary information to adjust for the misclassification of illicit drug use variables as follows:

As before, X represents one of the illicit drug use variables, which is assumed to have a two-part structure. To handle the missing values of X, we consider Xs the product of two ancillary variables: W a binary indicator of whether the subject was exposed, and Z, a continuous variable.

We used urine toxicology screen results to create an indicator variable I_1 as follows:

I = 1, if the urine toxicology screen result is positive for the illicit drug in question
= 0, otherwise.

To incorporate this supplementary information in the imputation scheme, we coded W and Z as follows:

W = 1, if X > 0
= 1, if I = 1 and X=0
= 1, if X is missing and I=1
= 0, if X=0 and I=0,

Z = X, if X > 0
= missing, if X=0
= missing, if X is missing.

By integrating the supplementary data into our motivating example, the binary indicator denoting subject exposure, W, became a fully observed variable. Subsequently, we applied MICE to impute the missing values of the log-transformed Z. For cases where W=0, the imputed exposure value was set to zero, while for W=1, the imputed value was obtained from exp(Z). We utilized custom scripts for data analysis, which are available on our GitHub repository https://github.com/takkaya/MICE-for-imputing-a-semi-continuous-covariate.git.

SIMULATION STUDY

Data Generation

We adopted the simulation scenarios outlined by.[18] The simulated datasets comprised a continuous exposure variable (T), a continuous outcome variable (Y), a semi-continuous confounder variable (X_1), a binary confounder variable (X_2) and a continuous confounder variable (X_3). We generated data in which 20%, 40%, 60% of the values of the semi-continuous variable X_1 were zero. The steps for generating the data are outlined below.

1. Two confounder variables (X_2, X_3) were as $X_2 \sim$ Binomial (1, 0.7) and $X_3 \sim$ N (0, 1) respectively.

2. A semi-continuous confounding variable (X_1) was simulated by first generating a binary indicator U, as a function of the confounder variables:

$$Pr(U=1|X_2,X_3)=expit(\theta_0+\theta_1X_2+\theta_2X_3) \quad (1)$$

with the value for θ_0 altered to control the percentage of zeros, and θ_1 and θ_2 are set to 1.25 and 0.40 respectively. A continuous variable (V), representing the positive values for X_1, was then generated from a Poisson regression model dependent on the auxiliary variables using $V \sim Poisson(\mu)$

$$\log(\mu)= \alpha_0 + \alpha_1X_2 + \alpha_2X_3 \quad (2)$$

where α_0=2.35, α_1=0.5, α_2=0.3. The semi-continuous variable, X_1, was then obtained by U * exp(log(V)).

3. A continuous exposure variable (T) was generated as a function of three confounder variables:

$$T_i= \gamma_0 + \gamma_1X_{i1} + \gamma_2X_{i2} + \gamma_3X_{i3} + \varepsilon_i^T \quad (3)$$

where $\varepsilon_i^T \sim (0, \sigma_e^2)$.

4. and a continuous outcome Y was generated as a function of the confounder variables and the continuous exposure variable.

$$Y= \beta_0 + \beta_1 * T + \beta_2 * X_1 + \beta_3 * X_2 + \beta_4 * X_3 + \varepsilon^Y \quad (4)$$

where $\varepsilon_i^Y \sim N(0, \sigma^2)$ and β_0=0.5, β_1=4, β_2=5.6, β_3=0.3 and β_4=0.8.

Generating Missing Data

We imposed the missing data under missing at random (MAR) mechanism where the missingness in X_1 was dependent on the confounder variable X_2 using the logistic regression model.

$$\log(X_1 \text{ is missing}) = \eta_0 + \eta_1 * X_2 \quad (5)$$

where η_1=1.9. The value of η_0 is controlled to have three scenarios where 20%, 36% of values in the semi-continuous confounder variable (X_1) was missing. For each scenario, 1000 datasets of 500 observations were generated. The sample size of 500 observations per dataset was chosen to be a realistic sample size for a cohort study and was motivated by the Detroit Longitudinal Cohort Study (which recruited n=480 in total, with n=337 included in the case study analysis).

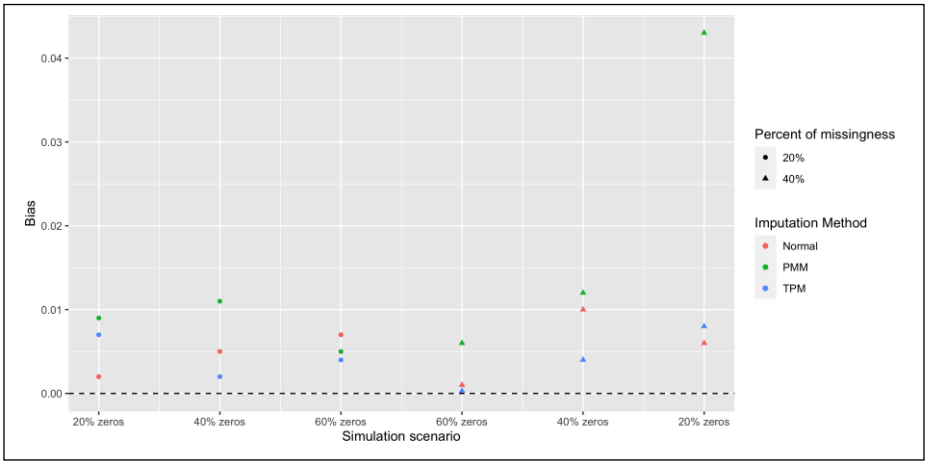
Estimation and Performance

The parameter of interest was the coefficient (β_1) from the linear regression for the continuous exposure variable on the continuous outcome (Equation 4). This parameter was estimated using three different imputation methods namely, multiple imputation of semi-continuous variables via two-part model, predictive mean matching and the imputation based on normal distribution. The "true" value was the parameter value specified in the data generating model (Equation 4). Four measures of performance were considered for the evaluation of the methods:

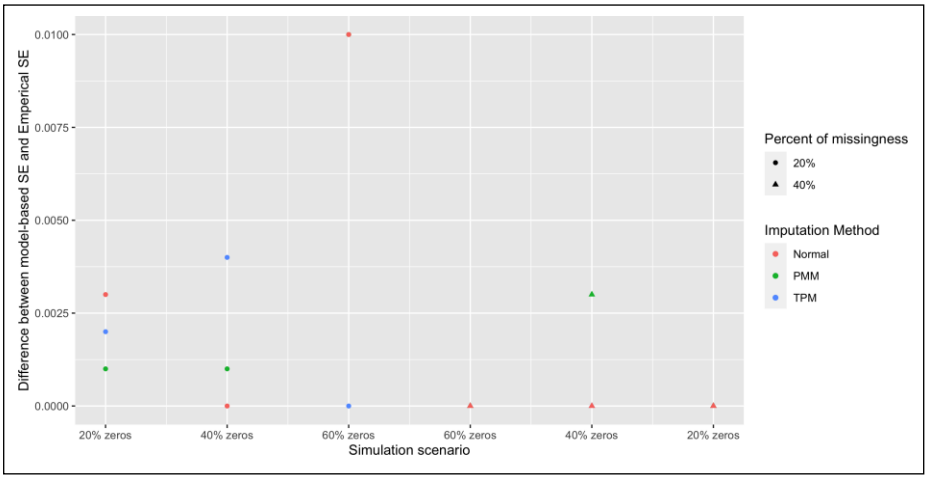
i. Bias: The difference between the average of the estimates (over the 1000 replications) and the "true" value.
ii. Empirical SE: The SD of the point estimates over the 1000 datasets.
iii. Model-based SE: The average of the estimated standard errors over the 1000 replications. If an imputation procedure is performing well, the average model-based SE should be like the empirical SE.
iv. Coverage: The proportion of 95% confidence intervals across the 1000 replications that contain the true value.

RESULTS

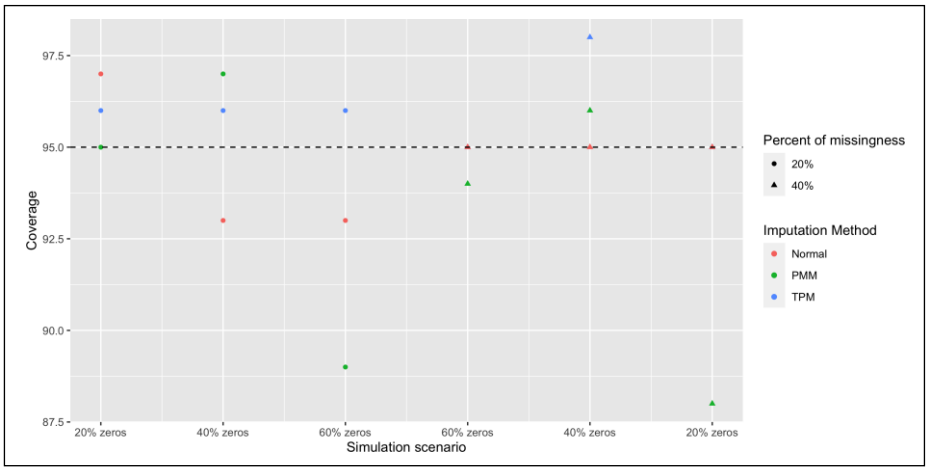
Results for the coefficient β_1 from the linear regression of the continuous exposure variable on the continuous outcome (Equation 4) are shown in Figure 1. As the percentage of zeros increases, imputation



(a)Empirical bias



(b) Difference between model-based SEs and empirical SEs



(c) Coverage of the estimates

PMM: Predictive mean matching; TPM: Two-part model; SE: Standard error.

FIGURE 1: Simulation results ((a) Empirical bias, (b) Difference between the model-based SEs and the empirical SEs, (c) Coverage of the estimates) for estimating the linear regression coefficient for the exposure variable in the presence of missing data in the semi-continuous covariate X_1.

based on the two-part model (TPM) produced the smallest bias across the scenarios we considered. Predictive mean matching also performed well; however, imputation based on TPM had over-coverage of the 95% confidence intervals in all scenarios. Although there were inconsistencies in scenarios with 20% missing observations, as the percentage of missing observations increases, the difference between model-based standard errors (SEs) and empirical SEs was smallest for the TPM method.

Overall, the TPM method performed reasonably well, particularly in scenarios where both the percentage of zeros and the percentage of missing observations in the semi-continuous covariate x_1 are relatively high. Predictive mean matching remains an acceptable approach for imputing semi-continuous covariates; how- ever, its performance deteriorates when the percentage of zeros approaches 60%.

DATA APPLICATION

We now demonstrate how we adapted the MICE method to handle missing data in the Detroit Longitudinal Study. This sample includes 480 pregnant women recruited at their first antenatal visit. Information on alcohol use was collected at each prenatal visit using a timeline follow-back interview.[19] These interview data were converted to ounces of absolute alcohol (AA) and summarized as average ounces of AA per day. Maternal reports of cocaine, marijuana, and opiate use (days per month) were obtained at every prenatal visit except the first. Additional data include maternal smoking during pregnancy (cigarettes per day), demographic background, number of pregnancies, maternal age at delivery, years of education, marital status, socioeconomic status depression (Beck Depression Inventory), intellectual competence (Peabody Picture Vocabulary Test; PPVT), stressful life events, and the Home Observation for Measurement of the Environment (HOME), a measure of the quality of intellectual stimulation provided by the parent(s).[20]

Our analysis focused on assessing the effect of prenatal alcohol exposure (AA/day) on the Freedom from Distractibility Index of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III), while accounting for a set of potential confounders. Out of 480 subjects, we included only the 336 individuals for whom the Freedom from Distractibility Index score was available. We used generalized propensity score methodology to adjust for these confounders when evaluating the association between prenatal alcohol exposure and the outcome of interest. However, missing values for several key control variables presented challenges. Table 1 shows the descriptive statistics and percentage of missing data for variables included in the propensity score estimation. Missing data was moderate-to-large for variables such as gestational age at initial screening; maternal depression scores at birth, 6 months, 12 months, and 7 years; HOME scores at 6 months and 7 years; biological mother's PPVT score; and prenatal exposure to cocaine, marijuana, and opiates. Missing rates for these variables ranged from 0.02% to 46.0%. If we used only complete cases to estimate the effect of prenatal alcohol exposure on WISC at age 7 years, we would be left with only 114 cases, eliminating 66% of the data. Discarding such many incomplete cases poses significant threats to the validity of the complete case analysis and reduces power to detect true effects.

Our analysis aimed to assess the effect of prenatal alcohol exposure (AA/day) on the Freedom from Distractibility Index of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III), adjusting for confounders. Of the 480 subjects, 336 had valid WISC-III scores. We used generalized propensity score methodology to control for confounders in evaluating the association between prenatal alcohol exposure and the outcome. Missing data on key control variables posed a challenge, with missing rates ranging from 0.02% to 46% for variables like maternal depression and prenatal drug exposure. Using complete cases would have reduced the sample to just 114 subjects, eliminating 66% of the data and risking reduced power and potential bias.

Another issue was the discrepancy between self-reported drug use and urine toxicology screens. Some subjects reported no drug use but had positive toxicology screens (19 for cocaine, 45 for marijuana, and 1 for opiates). Additionally, 12 subjects had no drug use data because they had only one prenatal visit. For these 12, we used retrospective drug use reports from a 13-month follow-up. We treated prenatal drug use as a

two-part variable, with each drug variable calculated as the product of a binary exposure indicator and the amount of exposure. Missing data on the continuous part was handled by setting it as missing for non-users, those with contradictory toxicology results, and those without prenatal data.

TABLE 1: The frequency distribution of potential confounders in the Detroit Longitudinal Cohort Study.

Continuous variables	**% of missing**	**Mean (SD)**
Socioeconomic status at birth/infancy	0.0	33.8 (30.9)
Socioeconomic status at age 7 yr follow-up visit	0.0	25.7 (11.1)
Gestational age at screening	2.0	23.7 (7.7)
Number of prenatal visits	0.0	5.18 (3.23)
Biological mother‘s education (yr)	0.0	11.7 (1.6)
Beck Depression Inventory score (Prenatal)	3.0	11.0 (7.6)
Beck Depression Inventory (6-mo postnatal visit)	10.0	11.4 (7.9)
Beck Depression Inventory (12-mo postnatal visit)	46.0	9.9 (7.1)
Beck Depression Inventory (7-yr follow-up)	0.2	8.1 (7.1)
Biological mother‘s verbal IQ (PPVT score)	9.0	72.1 (12.3)
Primary care giver‘s at 7-yr follow-up PPVT	0.0	73.6 (13.6)
Prenatal cocaine exposure	13.0	1.0 (2.9)
Prenatal marijuana exposure	13.0	0.8 (2.6)
Prenatal opiate exposure	13.0	0.2 (1.1)
Maternal smoking during pregnancy (cigarettes/day)	0.0	9.7 (11.7)
HOME score at birth/infancy	18.0	30.9 (4.7)
HOME score at 7-yr follow-up visit	0.5	39.5 (74.9)
Number of stressful events at 7-yr follow-up	0.0	9.8 (5.5)
Perceived life stress at 7-yr follow-up visit	0.0	34.6 (27.1)
Child’s age 7 yr follow-up visit (in days)	0.0	2830 (115.6)
Categorical variables	**% of missing**	**n (%)**
Biological mother’s marital status at recruitment		
Married	0.0	36 (11.0)
Not married	0.0	300 (89.0)
Primary care giver‘s marital status (7-yr follow-up visit)		
Married	0.0	287 (85.0)
Not married	0.0	49 (15.0)
Parity		
0	0.0	116 (34.5)
1	0.0	93 (27.7)
2	0.0	76 (22.6)
3	0.0	23 (6.8)
4	0.0	15 (4.5)
>=5	0.0	13 (3.9)
Gravidty		
1	0.0	55 (16.4)
2	0.0	76 (22.6)
3	0.0	68 (20.2)
4	0.0	41 (12.2)
>=5	0.0	96 (28.6)

SD: Standard deviation; PPVT: Peabody Picture Vocabulary Test; HOME: Home Observation for Measurement of the Environment.

We used the MICE package in R to impute missing values for control variables, generating 10 imputed datasets. For each dataset, we estimated the propensity score based on all key control variables and then fit an outcome model predicting the WISC-III Freedom from Distractibility Index as a function of prenatal alcohol exposure and the estimated propensity score. We combined the estimates and standard errors using Rubin's rules. A complete case analysis with 114 subjects revealed larger standard errors compared to the multiple imputation approach (Table 2). The relationship between the propensity score and the outcome differed significantly between the two methods, suggesting bias in the complete case analysis due to differences in covariance structures.

TABLE 2: Effect of prenatal alcohol exposure on the Wechsler Intelligence Scale for Children Freedom from Distractibility Index at age 7 years using complete cases and multiple imputation.

	Complete case			Imputation based on TPM		
	Estimate	**SE**	**95% CI**	**Estimate**	**SE**	**95% CI**
Intercept	104.6	2.0	(100.6, 108.6)	101.4	1.4	(98.6, 104.2)
Prenatal alcohol consumption	-8.5	6.6	(-21.8, 4.7)	-10.4	3.5	(-17.2, -3.6)
Propensity score	-7.9	10.5	(-28.7, 12.7)	2.0	8.4	(-14.5, 18.6)

TPM: Two-part model; CI: Confidence interval; SE: Standard error.

DISCUSSION

In this paper, we present a case study that demonstrates how the MICE package in R can be adapted to implement multiple imputation in settings where some incomplete covariates exhibit a semi-continuous structure. Buuren and Groothuis-Oudshoorn emphasize that the importance of employing tailored imputation models that align with the distributional properties of the data.[17]

Our multiple imputation strategy is straightforward to implement, allowing practitioners to use the MICE package with minimal coding. We evaluated and compared our method against one of the most used techniques, predictive mean matching. Research by Harel et al., explored various imputation strategies, including predictive mean matching and multiple imputation with different modeling assumptions.[21] Their findings indicate that while predictive mean matching is widely utilized, it may not always effectively capture the underlying data structure, particularly in scenarios with high missingness.

In our simulation study, we found that our approach performs well. Specifically, we observed that when both the percentage of missing observations and the percentage of zeros in the semi-continuous covariate are high, the imputation method based on the two-part model slightly outperforms other approaches. Although predictive mean matching yielded reasonable results in most cases, it does not accommodate supplementary information necessary to correct misclassification, as demonstrated in our motivating example. King et al. examined the type of missing data in substance use, highlighting the substantial bias occur when participants who are missing data are different from those who are not missing data.[22]

We illustrate our approach using data from the Detroit Longitudinal Cohort Study, which aims to assess the effects of prenatal alcohol exposure on child and adolescent development. A critical challenge in this analysis is the missing data on key control variables, including prenatal exposure to cocaine, marijuana, and opiates. These variables follow a two-part structure, where zero values indicate abstinence and non-zero values reflect varying degrees of exposure. This necessitates an imputation approach that preserves the distributional characteristics of the data.

Another complication arises from routine urine toxicology screening results, which indicated that some subjects reported no illicit drug use during pregnancy despite positive results for the drugs in ques-

tion. To address these mis-specified covariates, we extended our method to incorporate additional information from routine urine toxicology screen results and postnatal assessments. Our comparison of complete case analyses with results obtained from our imputation approach revealed significant differences between subjects with and without missing data, underscoring the substantial bias introduced by complete case analysis.

Looking ahead, our current work includes extending our proposed approach to situations where the exposure variable is also considered to have a semi-continuous structure. In the context of the Detroit study, prenatal alcohol exposure is an example of such a variable and adapting our approach to accommodate this would be straightforward, following the methodology described in reference 21.

Finally, we acknowledge that our work is limited by the simulation scenarios we considered, which focused on a partially observed semi-continuous covariate. Future research should conduct comprehensive simulation studies to evaluate the performance of our proposed imputation algorithm across multiple partially observed semi-continuous covariates, thereby enhancing its applicability and robustness in diverse settings.

CONCLUSION

In conclusion, our study introduces a novel approach to addressing missing semi-continuous covariate data in propensity score estimation by adapting the R package MICE and incorporating misclassification corrections. This methodology demonstrated significant improvements in imputing maternal substance use variables, such as prenatal alcohol and illicit drug exposure, which often exhibit a zero-inflated and long-tailed distribution. Specifically, our two-part imputation model more accurately handled covariates with high proportions of zeros, compared to traditional methods like predictive mean matching.

For prenatal alcohol exposure (AA/day), the proposed method reduced bias in propensity score estimation by accounting for discrepancies between maternal self-reports and biological assay results, resulting in more balanced covariates across exposure groups. Similarly, the correction of misclassified illicit drug use data significantly decreased the underestimation of exposure effects on childhood cognition.

These enhancements led to more precise estimates of the relationship between prenatal substance exposure and cognitive outcomes, as measured by the Wechsler Intelligence Scale for Children-Third Edition (WISC-III). Notably, the two-part imputation model was particularly effective for variables with >30% missing observations and >40% zero values, demonstrating superior performance over conventional imputation techniques.

Future research can expand this framework to examine additional semi-continuous variables in observational studies, addressing a broader range of public health challenges related to maternal behaviors and child development. This work underscores the importance of aligning imputation strategies with the unique distributional properties of the data, ultimately enhancing the validity and reliability of causal inferences in environmental epidemiology.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Tuğba Akkaya Hocagil, Louise M. Ryan;* ***Design:*** *Tuğba Akkaya Hocagil, Richard J. Cook, Louise M. Ryan;* ***Control/Supervision:*** *Louise M. Ryan, Richard J. Cook;* ***Data Collection and/or Processing:*** *Sandra W. Jacobson, Joseph Jacobson;* ***Analysis and/or Interpretation:*** *Tuğba Akkaya Hocagil, Sandra W. Jacobson, Joseph Jacobson;* ***Literature Review:*** *Tuğba Akkaya Hocagil;* ***Writing the Article:*** *Tuğba Akkaya Hocagil, Louise M. Ryan;* ***Critical Review:*** *Richard J. Cook, Sandra W. Jacobson, Joseph Jacobson;* ***References and Fundings:*** *Joseph Jacobson, Sandra W. Jacobson;* ***Materials:*** *Joseph Jacobson, Sandra W. Jacobson.*

REFERENCES

1. Imai K, van Dyk DA. Causal Inference with General Treatment Regimes. J Am Stat Assoc. 2004;99(467):854-66. [Crossref]
2. Akkaya Hocagil T, Cook RJ, Jacobson SW, Jacobson JL, Ryan LM. Propensity Score Analysis for a Semi-Continuous Exposure Variable: A Study of Gestational Alcohol Exposure and Childhood Cognition. J R Stat Soc Ser A Stat Soc. 2021;184(4):1390-413. [Crossref] [PubMed] [PMC]
3. Rubin DB, Thomas N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. Biometrics. 1996;52(1):249. [Crossref]
4. Rubin DB. Inference and Missing Data. Biometrika. 1976;63(3):581. [Crossref]
5. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Wiley; 2002. [Crossref]
6. Carpenter JR, Kenward MG. Multiple Imputation and its Application. Wiley; 2013. [Crossref]
7. Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J, et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res. 2019;28(1):3-19. [Crossref] [PubMed] [PMC]
8. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res. 2016;25(1):188-204. [Crossref] [PubMed]
9. Coffman DL, Zhou J, Cai X. Comparison of methods for handling covariate missingness in propensity score estimation with a binary exposure. BMC Med Res Methodol. 2020;20(1):168. [Crossref] [PubMed] [PMC]
10. Vink G, Frank LE, Pannekoek J, van Buuren S. Predictive mean matching imputation of semicontinuous variables. Stat Neerl. 2014;68(1):61-90. [Crossref]
11. Van Buuren S. Multiple imputation of multilevel data. The Handbook of Advanced Multilevel Analysis; 2011. p.173-96.
12. White IR, Wood AM. Tutorial in Biostatistics Multiple imputation using chained equations: Issues and guidance for practice. 2011; (July 2010). [Crossref] [PubMed]
13. Lee KJ, Carlin JB. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. Am J Epidemiol. 2010;171(5):624-32. [Crossref] [PubMed]
14. Burton A, Billingham LJ, Bryan S. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. Clinical Trials. 2007;4(2):154-61. [Crossref] [PubMed]
15. Su YS, Gelman A, Hill J, Yajima M. Multiple Imputation with Diagnostics (mi) in R : Opening Windows into the Black Box. J Stat Softw. 2011;45(2). [Crossref]
16. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley; 1987. [Crossref]
17. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3). [Crossref]
18. Nguyen CD, Moreno-Betancur M, Rodwell L, Romaniuk H, Carlin JB, Lee KJ. Multiple imputation of semi-continuous exposure variables that are categorized for analysis. Stat Med. 2021;40(27):6093-106. [Crossref] [PubMed]
19. Jacobson SW, Chiodo LM, Sokol RJ, Jacobson JL. Validity of Maternal Report of Prenatal Alcohol, Cocaine, and Smoking in Relation to Neurobehavioral Outcome. Pediatrics. 2002;109(5):815-25. [Crossref] [PubMed]
20. Hollingshead L, Childs RA. Reporting the Percentage of Students above a Cut Score: The Effect of Group Size. Educational Measurement: Issues and Practice. 2011;30(1):36-43. [Crossref]
21. Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):576-84. [Crossref] [PubMed] [PMC]
22. King C, Englander H, Priest KC, Korthuis PT, McPherson S. Addressing Missing Data in Substance Use Research: A Review and Data Justice-based Approach. J Addict Med. 2020;14(6):454-6. [Crossref] [PubMed] [PMC]