Yang *et al.*
BMC Medical Research Methodology (2023) 23:41
https://doi.org/10.1186/s12874-023-01847-2

BMC Medical Research Methodology

RESEARCH

Open Access

Propensity score analysis with missing data using a multi-task neural network

Shu Yang[1†], Peipei Du[2,3†], Xixi Feng[4†], Daihai He[3], Yaolong Chen[5], Linda L. D. Zhong[6,7], Xiaodong Yan[8*] and Jiawei Luo[2*]

Abstract

Background Propensity score analysis is increasingly used to control for confounding factors in observational studies. Unfortunately, unavoidable missing values make estimating propensity scores extremely challenging. We propose a new method for estimating propensity scores in data with missing values.

Materials and methods Both simulated and real-world datasets are used in our experiments. The simulated datasets were constructed under 2 scenarios, the presence (T = 1) and the absence (T = 0) of the true effect. The real-world dataset comes from LaLonde's employment training program. We construct missing data with varying degrees of missing rates under three missing mechanisms: MAR, MCAR, and MNAR. Then we compare MTNN with 2 other traditional methods in different scenarios. The experiments in each scenario were repeated 20,000 times. Our code is publicly available at https://github.com/ljwa2323/MTNN.

Results Under the three missing mechanisms of MAR, MCAR and MNAR, the RMSE between the effect and the true effect estimated by our proposed method is the smallest in simulations and in real-world data. Furthermore, the standard deviation of the effect estimated by our method is the smallest. In situations where the missing rate is low, the estimation of our method is more accurate.

Conclusions MTNN can perform propensity score estimation and missing value filling at the same time through shared hidden layers and joint learning, which solves the dilemma of traditional methods and is very suitable for estimating true effects in samples with missing values. The method is expected to be broadly generalized and applied to real-world observational studies.

Keywords Observational study, Propensity score analysis, Neural network, Multitasking learning, Causal effect estimation, Inverse probability weighting

†Shu Yang, Peipei Du and Xixi Feng contributed equally to this work.

*Correspondence:
Xiaodong Yan
yanxiaodong@sdu.edu.cn
Jiawei Luo
2111952576@qq.com
[1] School of Intelligent Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu, China
[2] West China Biomedical Big Data Center, West China Hospital/West China School of Medicine, Sichuan University, Chengdu, China
[3] Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China
[4] School of Public Health, Chengdu Medical College, Chengdu, China
[5] Institute of Health Data Science, Lanzhou University, Lanzhou, China
[6] Biomedical Sciences and Chinese Medicine, School of Biological Sciences, Nanyang Technological University, Singapore, Singapore
[7] School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, China
[8] School of Economics, Shandong University, Jinan, China

Introduction

In observational studies, propensity scores are increasingly used to control for confounding [1, 2]. When the observed baseline characteristics are sufficient to correct for confounding bias and the propensity model is correctly constructed, a conditional exchange can be conducted between subjects with the same propensity score [3, 4]. Observational studies usually inevitably have covariates with missing values. Currently, estimating the propensity score in the presence of missing values is a challenge for studying causality [5–8]. Common approaches to dealing with missing values in propensity analysis include full-case analysis, adding missing indicator variables to the propensity model, and multiple imputation [9–11]. Unfortunately, these methods are inherently flawed. For example, the missing indicator method introduces new biases [12]. There are studies using machine learning methods to replace traditional logistic regression [13–17]. However, they do not address the propensity score misestimation problem caused by overfitting. In contrast to hand-crafted models [18], neural networks can automatically learn interactions between variables. A multi-task neural network is a network structure with multiple outputs. It has been widely used in the medical field. With a multi-task neural network, propensity score computation and missing value filling can be performed jointly. By optimizing the global objective function, overfitting to the propensity score calculation task can be prevented, while the estimation problem of missing value [19] is effectively solved. This study develops a new pipeline for calculating propensity scores in samples with missing values based on a multi-task neural network. To evaluate the accuracy of our model in estimating the true effect, we conduct experiments on simulated and real-world data separately, and compare our method with traditional methods.

Data and methods

Propensity score

In a study, individual subjects may have multiple covariates. Propensity scoring is a way of simplification multiple covariates [20]. It condenses multiple covariates into a single variable (propensity score), whose meaning is the conditional probability of being assigned to the experimental group depending on the covariates [21]. A propensity score can be viewed as a function of the original multiple covariates, so the propensity score includes information about these covariates. Rosenbaum and Rubin demonstrated that the propensity score $e(X)$ can be used to balance the distribution of a covariate between experimental and control groups when the covariate X meets the strong negligibility assumption [3].

$$e(X_i) = \Pr\left(T_i = 1 | X_i\right)$$

Propensity score estimation

In complete data, logistic regression is the most commonly used method for estimating propensity scores under the conditions of binary treatment or exposure [22]. The propensity score is calculated by performing binary regression on covariates (i.e. potential confounders) by treatment or exposure indicator variables, which can be written as:

$$\mathrm{logit}(\mathrm{p_i}(\mathrm{T} = 1)) = \mathrm{X}'\beta, \mathrm{i} = 1, 2, \ldots, \mathrm{n}$$

where, $X' = (1, X_1, X_2, \ldots, X_K)$, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_K)'$, K is the number of covariates and n is the number of observations. An individual's propensity score can be estimated as

$$p_i = \frac{\mathrm{e}^{X_i'\beta}}{1 + \mathrm{e}^{X_i'\beta}}$$

In many situations, logistic regression may not be the best choice when estimating propensity scores. We assumed that the log probability of exposure was linearly related to covariates when using logistic regression to estimate exposure probabilities. However, this assumption is not always true. Logistic regression cannot estimate propensity scores accurately when covariates interact with each other or when covariates and treatments are not linear. To solve the inherent problem of logistic regression estimation of propensity scores, some studies substitute machine learning algorithms for logistic regression. These include decision trees, random forests, Naive Bayes, support vector machines, etc. [13–15, 23, 24] It is claimed that these methods can provide a more accurate estimate of propensity scores. Nevertheless, these conclusions have not been validated by systematic simulation studies.

Missing data

In realistic observational studies, individual covariates may have large amounts of missing data, which may lead to both loss of efficiency and biased estimates. Based on the degree to which confounding factors are related to outcome and exposure, the magnitude of bias varies.

Type of missing data

There are three types of missing data depending on the mechanism of missing: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [25, 26]. MCAR refers to missing data when a random subset of the study population has the same probability of being missing. In contrast to MCAR,

the term MAR is counterintuitive. MAR occurs when the probability of missing is dependent only on the observed information. Missing data are denoted by MNAR when their probability depends on the unobserved data, such as the observation value itself.

Methods for handling missing values

Complete case analysis is the easiest way to deal with incomplete confounding data, which restricts the analysis to cases where all variables are complete. If the absence of covariates is independent of treatment and outcome, then this approach provides unbiased estimates of group effects. Another simple method is the missing indicator method [27]. Before incorporating confounding into a propensity score model, add a "missing" category to partially observed categories. Continuous confounders are set to a specific value, such as 0, and both the confounding factor and missingness indicator (a variable that indicates whether the variable is observed) are included in the propensity score model. In many cases, this approach leads to biased results. Missing pattern analysis is a generalization of the missing index method. This method is used when all individuals are grouped together according to different missing patterns. Then, propensity scores are estimated in each group separately. As a practical matter, this method fails when the number of participants with missing patterns is lower than the number of observed covariates. It usually occurs when there are a lot of missing patterns in the data. Multiple imputation is a method of using chain equations to impute incomplete data, in which the missing covariates are imputed with plausible values based on the predicted distribution of the missing covariates in a set of observed data many times to create complete datasets [28, 29]. We used MICE (version 3.3.0) in R (version 3.6.3) to perform multiple imputation. A Bayesian linear regression was used for the mice model. It is commonly used when covariates and outcomes are continuous. Other parameters are set as defaults.

Inverse probability weighting

Inverse probability weighting (IPW) uses the inverse of the propensity score as weights to create a synthetic sample in which the baseline covariate distribution is independent of treatment assignment [30]. In this study, we use IPW to estimate the true effect. Unlike propensity score matching, IPW uses all individuals in both groups, thus avoiding sample waste. A high level of statistical power was maintained in all cases to detect effects. IPW was more sensitive to erroneous propensity score estimation. This limitation emphasizes the importance of carefully defining model selection before applying propensity score weighting. Multi-task neural networks can overcome this limitation.

Multi-task neural network

Neuronal networks are excellent function approximators, which can estimate linear and nonlinear functions. It uses data samples with known outcomes as examples for supervised training. In this process, a nonlinear function model is built to predict the output data based on the input data. Figure 1 **(a)** shows three independent neural networks. All networks have the same inputs and outputs. Back-propagation is used to train each net separately. There is no connection between the three nets, so the information that one learns cannot help the others. This is known as single-task learning (STL). Figure 1 **(b)** shows a single net with the same inputs as those on the left, but three outputs corresponding to the learning task. Each of the 3 outputs is connected to the same hidden layer. Three of the MTL outputs undergo parallel back-propagation. These results share a hidden layer, meaning the internal representation of one task is available for other tasks. The core idea of multitask learning is to share knowledge learned from different tasks and to train them simultaneously.

In this study, we propose a novel pipeline using a multi-task neural network (MTNN) to estimate propensity scores. There are three parts to our task set: reconstructing input covariates, estimating propensity scores, and predicting missing patterns. There is a close relationship between these tasks. The structure of MTNN is shown in Fig. 1. In order to achieve joint optimality across all tasks, the MTNN must correctly learn the relationship between covariates, covariates and absence, and covariates and exposure levels. Through joint learning and sharing hidden layers, MTNN reduces overfitting when estimating propensity scores. The detailed calculation procedure and more information about MTNN training can be found in Supplementary S1. Our tutorial and source code for MTNN are also available on github[1] so readers can apply our method to real problems and gain a deeper understanding of it. Models for missing value imputation and estimation of propensity scores are determined from the convergence of the objective function. In all experiments in this study, we chose the model for the last epoch after convergence.

Data

Simulation data

We adopted a data simulation generation process similar to that of Choi [7]. Two scenarios were considered, one in which the outcome was treatment-related (effect$\neq$0), and one in which it was treatment-independent (effect$=$0). In each scenario, we considered three different deletion

[1] https://github.com/ljwa2323/MTNN.

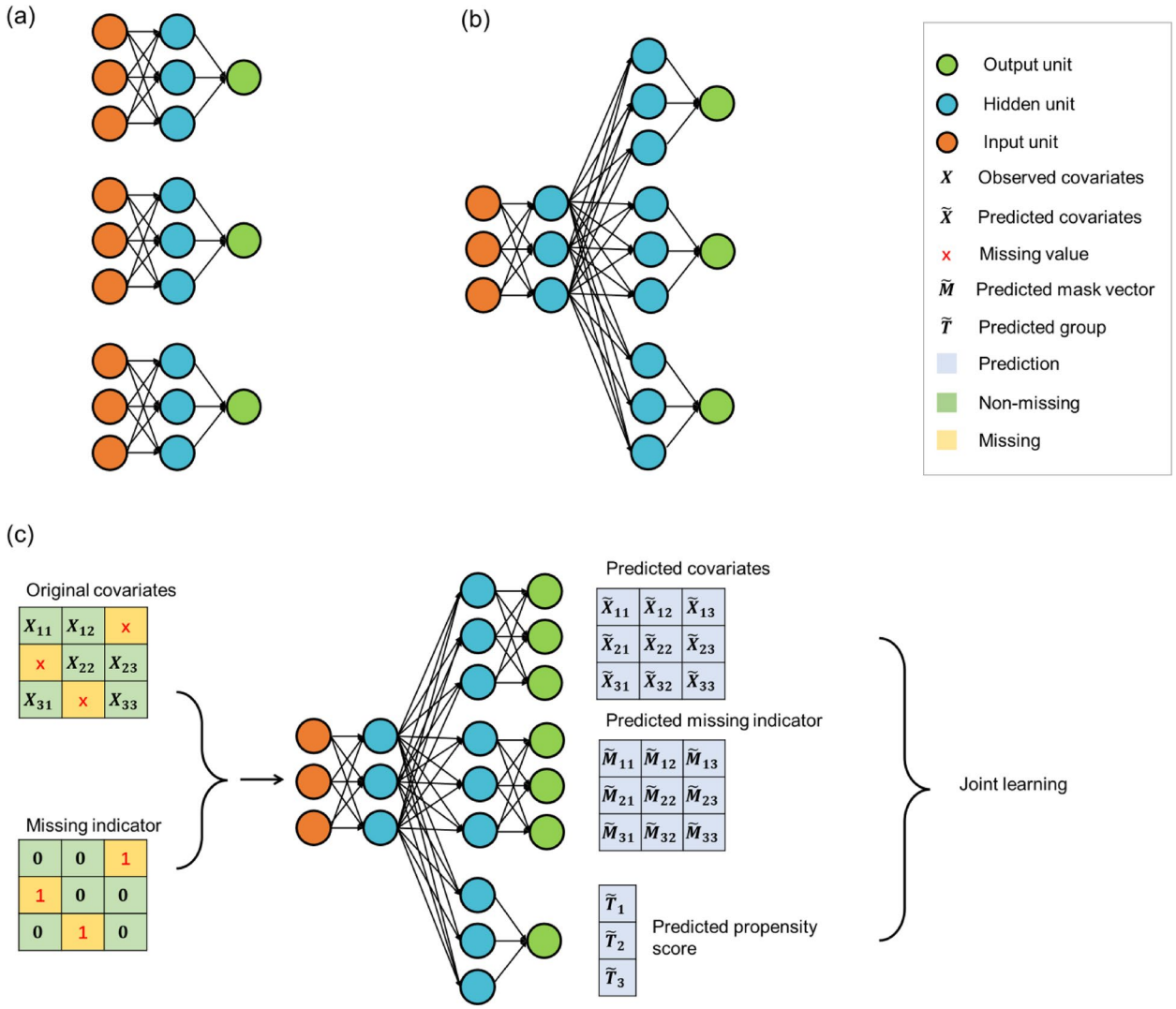

Fig. 1 Structure diagram of multi-task neural network

mechanisms. First, we generated 2 continuous covariates, X_1 and X_2, for each subject. X_1 follows a normal distribution with mean 0 and standard deviation 1. X_2 depends on X_1.

$$X_{2i} = 0.5X_{1i} + \varepsilon_i \text{ with } \varepsilon_{\mathrm{i}} \sim \mathrm{N}(0, 0.75)$$

In this way, the standard deviation of X_2 is also 1, and the correlation between X_1 and X_2 is equal to 0.5. The treatment T was generated from the binomial distribution, with the probability for subject I to receive the treatment being equal to:

$$\mathrm{logit}(\mathrm{P}(\mathrm{T_i} = 1|\mathrm{X_{1i}}, \mathrm{X_{2i}})) = -0.8 + 0.5\mathrm{X_{1i}} + 0.5\mathrm{X_{2i}}$$

By this equation, about 30% of subjects were treated. We constructed 2 scenarios:

Scenario 1: the outcome is affected by treatment: we assume, without losing generality, that treatment has an effect of 1 on the subject's outcome.

$$Y_i = X_{1i} + X_{2i} + Treat_i + \varepsilon_i, \text{with } \varepsilon_{\mathrm{i}} \sim \mathrm{N}(0, 1)$$

Scenario 2: the outcome is unrelated to the treatment.

$$Y_i = X_{1i} + X_{2i} + \varepsilon_i, \text{with } \varepsilon_{\mathrm{i}} \sim \mathrm{N}(0, 1)$$

To test the effect of different missing rates on effect estimation in simulated datasets, we preset 7 missing rates, including 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. Missing values in X_2 are generated using three mechanisms:

(1) MCAR: In X_2, a random proportion of observations is set to be missing.

(2) MAR: The higher the value of X_1, the more likely the value of X_2 is missing. Taking M as the missing indicator of X_2, the probability of missing X_2 value is:

$$\text{logit}(\text{P}(\text{M}_\text{i} = 1)) = \text{X}_{1\text{i}} + \text{C}$$

(3) MNAR: The higher the value of X_2, the more likely the value is missing. The probability of missing an X_2 value is:

$$\text{logit}(\text{P}(\text{M}_\text{i} = 1)) = \text{X}_{2\text{i}} + \text{C}$$

C is a constant used to control the missing rate. As an example, if a missing rate of around 50% is to be controlled, C can be set to 0.

Real-world data

The real-world data come from a subset of the data from the treated group in the National Supported Work Demonstration (NSWD) and the comparison sample from the Population Survey of Income Dynamics (PSID). The dataset has been used by many researchers to test the effects of different propensity score analysis methods [31, 32]. There are 614 samples in this dataset (185 treatments and 429 controls). Each person has 9 variables. Table S1 provides more details. Treat is the intervention variable, re78 is the outcome, and the other 7 variables are covariates. Table S2 summarizes the distribution of covariates between different treatment groups. It shows that the distributions of the variables age, race, married, nondegree, re74, re75 differ between groups. Therefore, we need to correct the effect estimates with propensity scores.

Our experiments used the inverse probability-weighted effect size of the propensity score calculated from the complete data as the reference. Simulations were then performed to estimate the true effect under the three missing mechanisms. We made missing values occur in both variables re74 and re75. In each of these variables, missing values were constructed randomly. Similar to the setting we used for simulated datasets, we used 7 missing rate settings for real-world datasets: 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8.

(1) MCAR: In both variables re74 and re75, randomly selected given proportion of observations are set to be missing.

(2) MAR: The missing rate is assumed to be proportional to a linear combination of age and education. These 2 variables were chosen arbitrarily without loss of generality, as there were correlations between the covariates (table S8). To facilitate setting the probability of missing, we normalize the age and years of education so that the mean is 0. Let M_1 and M_2 represent the missing indicators of re74 and re75, respectively, then their missing probability is:

$$\text{logit}(\text{P}(\text{M}_{\text{i}1} = 1)) = \text{age}_\text{i} + \text{educ}_\text{i} + \text{C}$$

$$\text{logit}(\text{P}(\text{M}_{\text{i}2} = 1)) = \text{age}_\text{i} + \text{educ}_\text{i} + \text{C}$$

(3) MNAR: The higher the value of a variable, the more likely that value is missing. Similar to age and years of education, we also normalize re74 and re75. Then the probability of re74/re75 missing is:

$$\text{logit}(\text{P}(\text{M}_{\text{i}1} = 1)) = \text{re74}_\text{i} + \text{C}$$

$$\text{logit}(\text{P}(\text{M}_{\text{i}2} = 1)) = \text{re75}_\text{i} + \text{C}$$

Estimation of the true effect

The first step is to deal with missing values in the samples. As MTNN computes propensity scores and imputation values simultaneously, it does not require separate missing value processing. When propensity scores were estimated by logistic regression, multiple imputation and missing indicator methods were used to handle missing values. We estimate propensity scores using age, education, race, marital status, education, and re74 and re75 as covariates. These 7 covariates are also included in the regression analysis used to estimate effect. Lastly, we estimated the effect using an inverse probability-weighted regression analysis of the propensity score, in which subjects receiving treatment were weighed 1/propensity score and subjects not receiving treatment were weighed 1/(1 - propensity score). Figure 2 shows the workflow for estimating the effects of the three methods.

Evaluation

There are 2 kinds of effects in the experiments with simulated data, and three mechanisms for handling missing values, i.e., 6 scenarios for generating simulated data, and 3 methods for handling missing values. In experiments with real-world data, there are three missing mechanisms, namely three scenarios. For each scenario, the same process of missing value imputation, propensity score calculation, and effect estimation was repeated 20,000 times before evaluating the results of the different methods. Comparisons are conducted based on standard deviations (SD) and root mean square errors (RMSE), which is defined as:

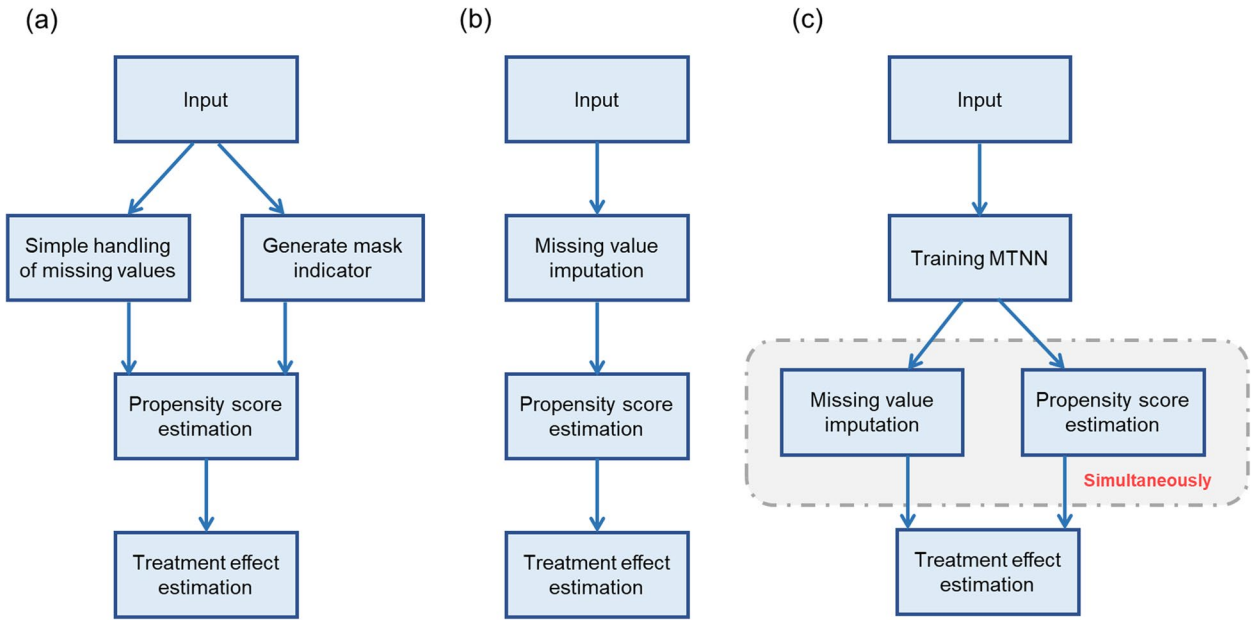

Fig. 2 Flowchart of the three methods for estimating effect. a the missing index method; b the multiple imputation method; c multi-task neural network method. MTNN, multi-task neural network

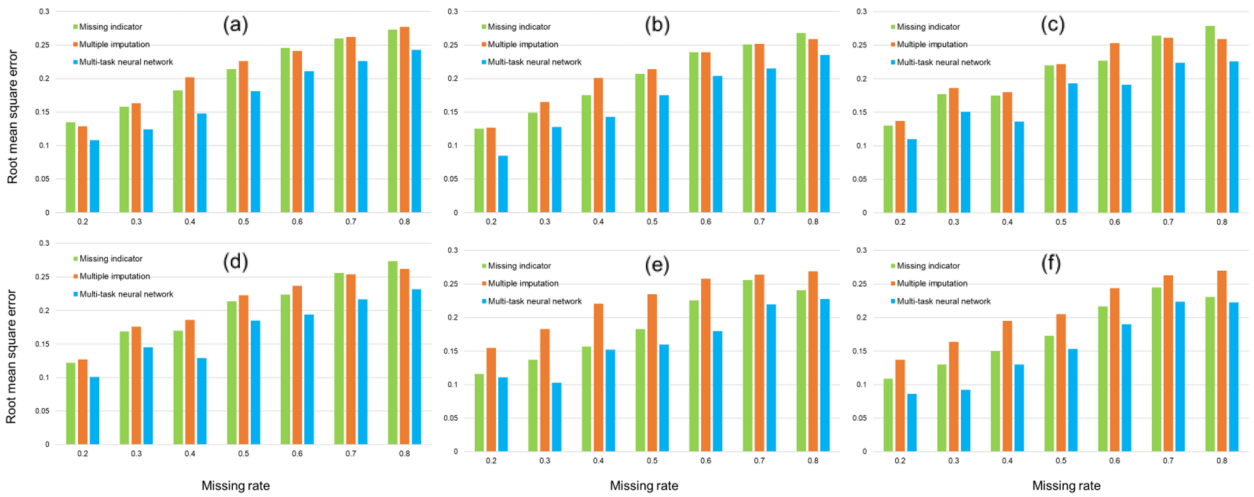

Fig. 3 Root mean square error of the true effect estimated by different methods under three missing mechanisms in the simulation dataset. (a), (d) are under MCAR, (b), (e) are under MCAR, (c)-(f) is under MNAR. For **a**, **b** and **c**, the true effect is 0; for **d**, **e** and **f**, the true effect is 1. MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_i - \beta\right)^2}$$

Where $\hat{\beta}$ is the estimate and β is the true value.

Result

Analysis results on simulation datasets

Figure 3 shows the RMSE of the true effect estimates under 2 true effect scenarios and three missing

mechanisms. The smallest RMSE for all 6 data scenarios is achieved with MTNN. Thus, MTNN seems to be the best method over the other two. In addition, regardless of the choice of method used, the higher the missing rate, the higher the RMSE. When the missing rate was increased from 0.2 to 0.8, the RMSE for any of the three estimation methods nearly doubled. Table 1, Table S3 and Table S4 present more detailed information on the estimation results for the three methods. In all scenarios of data, we find that MTNN is not only optimal in estimation of true effect deviation, but also that the standard deviation of its estimation results is the smallest. This shows that MTNN provides the most accurate estimation, as well as being more stable than other methods.

Analysis results on real-world datasets

We first calculated the propensity score by logistic regression from the complete data, and then used the inverse probability-weighted regression equation to calculate the effect to be 712.743 (Table S7). Since the true effect of real-world data is unknowable, we use it as a reference standard to compare the performance of different methods.

Figure 4 compares RMSE between different methods under three distinct missing mechanisms. According to

Table 1 Estimation of the true effect in the simulated datasets using three different methods under the MCAR mechanism

Missing rate	Method	True effect =0			True effect =1		
		Mean	SD	RMSE	Mean	SD	RMSE
0.2	Missing indicator	0.119	0.059	0.131	1.119	0.059	1.12
	Multiple imputation	0.11	0.071	0.128	1.11	0.071	1.112
	Multi-task neural network	0.091	0.052	0.103	1.077	0.055	1.079
0.3	Missing indicator	0.146	0.063	0.157	1.146	0.063	1.147
	Multiple imputation	0.136	0.076	0.153	1.136	0.076	1.138
	Multi-task neural network	0.107	0.057	0.12	1.12	0.052	1.121
0.4	Missing indicator	0.173	0.061	0.182	1.173	0.061	1.174
	Multiple imputation	0.192	0.089	0.209	1.192	0.089	1.195
	Multi-task neural network	0.138	0.05	0.146	1.139	0.058	1.141
0.5	Missing indicator	0.206	0.069	0.216	1.206	0.069	1.207
	Multiple imputation	0.214	0.078	0.226	1.214	0.078	1.216
	Multi-task neural network	0.173	0.058	0.181	1.169	0.06	1.17
0.6	Missing indicator	0.234	0.071	0.243	1.234	0.071	1.236
	Multiple imputation	0.228	0.076	0.239	1.228	0.076	1.23
	Multi-task neural network	0.2	0.064	0.208	1.198	0.06	1.199
0.7	Missing indicator	0.242	0.081	0.254	1.242	0.081	1.244
	Multiple imputation	0.248	0.08	0.259	1.248	0.08	1.251
	Multi-task neural network	0.207	0.078	0.219	1.204	0.071	1.206
0.8	Missing indicator	0.258	0.061	0.264	1.258	0.061	1.259
	Multiple imputation	0.26	0.074	0.269	1.26	0.074	1.262
	Multi-task neural network	0.226	0.053	0.232	1.224	0.057	1.225

SD, standard deviation; RMSE, root mean square error; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random

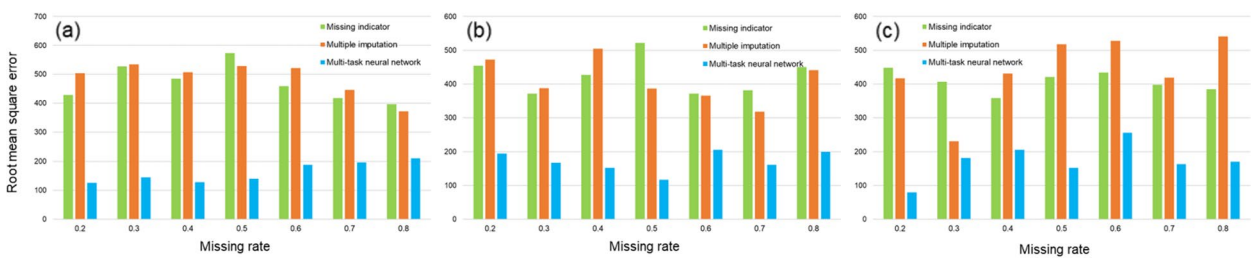

Fig. 4 RMSE of the true effect estimated by different methods under three missing mechanisms in the real-world dataset. **a** MCAR, **b** MAR, **c** MNAR

Table 2 Estimation of the true effect in the real-world datasets using three different methods under the MCAR mechanism

Missing rate	Method	Mean	SD	RMSE
0.2	Missing indicator	352.262	250.792	431.108
	Multiple imputation	198.329	276.932	576.882
	Multi-task neural network	672.620	121.163	121.075
0.3	Missing indicator	395.330	283.958	415.239
	Multiple imputation	403.878	309.952	425.198
	Multi-task neural network	718.292	144.233	136.098
0.4	Missing indicator	316.312	226.881	450.458
	Multiple imputation	277.395	247.175	493.797
	Multi-task neural network	736.417	146.882	140.491
0.5	Missing indicator	240.517	266.086	534.726
	Multiple imputation	341.053	279.437	455.591
	Multi-task neural network	683.922	112.963	110.333
0.6	Missing indicator	339.664	233.459	433.169
	Multiple imputation	171.271	191.980	570.923
	Multi-task neural network	623.39	140.996	160.172
0.7	Missing indicator	323.533	202.261	433.415
	Multiple imputation	318.219	232.415	451.292
	Multi-task neural network	587.779	116.185	166.178
0.8	Missing indicator	328.838	128.502	402.568
	Multiple imputation	395.226	123.355	338.146
	Multi-task neural network	640.485	167.939	174.043

SD, standard deviation; RMSE, root mean square error; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random

the analysis results of simulated data, MTNN exhibited the smallest RMSE under different missing mechanisms and missing rates. The difference is that in the real-world dataset, the missing rate is less influential on the RMSE of the estimated result. Table 2, Table S5 and Table S6 provide further details of the estimation results for the various methods. It is clear that the standard deviation of the MTNN estimation results is lower than that of the 2 other methods. Figures 5, 6 and 7 show the between-group standardized mean differences (SMD) of each covariate adjusted by the propensity scores estimated by the three methods under the three missing mechanisms.

Discussion

In this study, we develop a novel method for calculating propensity scores with multi-task neural networks that can calculate propensity scores directly for samples with missing values. On simulated and real-world datasets, we compare the proposed method with two commonly used ones. Under the three missing mechanisms, the RMSE of our proposed method for estimating the true effect is the smallest. In addition, the standard deviation of the true effect estimated by MTNN is the smallest, indicating that it is more robust than the other two methods. While previous studies have demonstrated smaller RMSEs for machine learning algorithms, our study confirms these findings in scenarios with missing values [33–36]. We also found that under lower missing rate conditions, the RMSE of the missing indicator method is better than multiple imputation for all 3 missing mechanisms. This result is consistent with the previous study [7].

Recent studies have used autoencoders to reduce the dimension of high-dimensional features and then calculate propensity scores using the reduced features [17]. It leverages the ability of neural networks to deal with high-dimensional data. However, they did not consider reconstruction and computation of the propensity score as joint tasks. Instead, we train the model together with reconstruction of the input, prediction of missing patterns, and estimation of propensity scores as joint tasks to prevent overfitting. It causes propensity scores to be close to zero or one, resulting in biased estimates of the effects.

As the variable dimension increases in observational studies, the relationship between variables will be more complex, and missing will be more difficult to avoid. It also becomes increasingly difficult to manually determine propensity models for high-dimensional variables. The neural network has the ability to model complex models, so there is no need to manually specify the so-called correct model, and the neural network can learn adaptively by observing the data. Multiple imputation is expensive for large datasets. In contrast, for the MTNN model, the computational cost of this process is smaller. Furthermore, Compared to multiple imputation [37], MTNN does not require any prior assumptions about the distribution of the data. It automatically learns the correlations between variables, thus impute their missing values.

In practice, a missing rate of greater than 30% is generally considered too high to make a reliable inference, but we want to thoroughly test the MTNN model's stability and performance under different missing rate scenarios. Due to this, we have created a list of missing rates that are relatively high. We found that even when the missing rate is high, MTNN still performs well. It shows that the correlation between variables can be captured and utilized very effectively. Even though an increase in missing rates decreases the performance of the MTNN model, it still outperforms other methods.

Limitations

Our study also has some limitations. First, there is a slight difference in performance between simulated and real data for the MTNN model. The reason for this phenomenon is that in real-world data, relationships between variables are more complex. It is difficult to simulate these unknowable complex connections manually. Due to the fact that our experiments simulate only the simplest

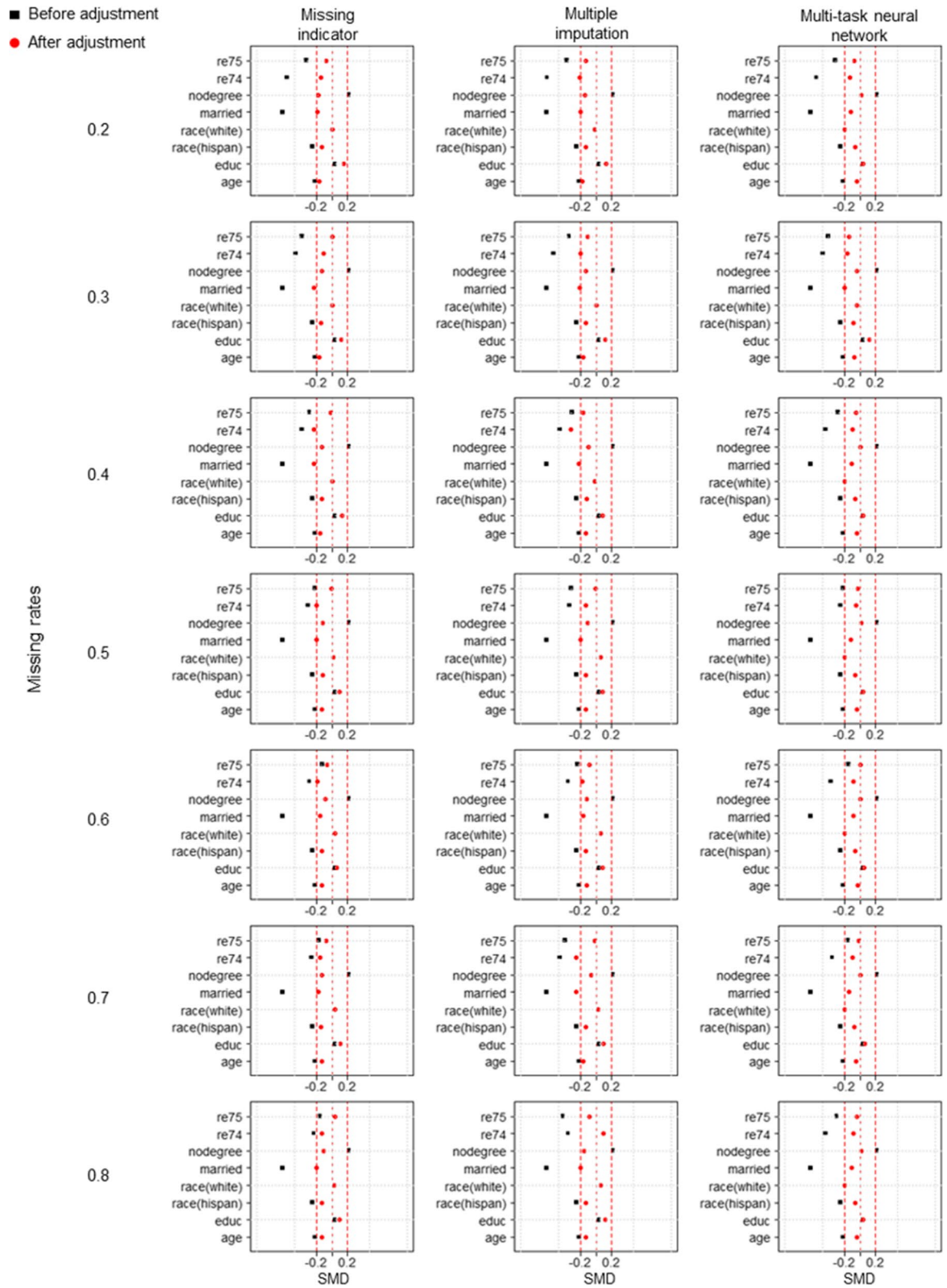

Fig. 5 Between-group standardized mean differences under MCAR for covariates adjusted for propensity scores calculated by three different methods

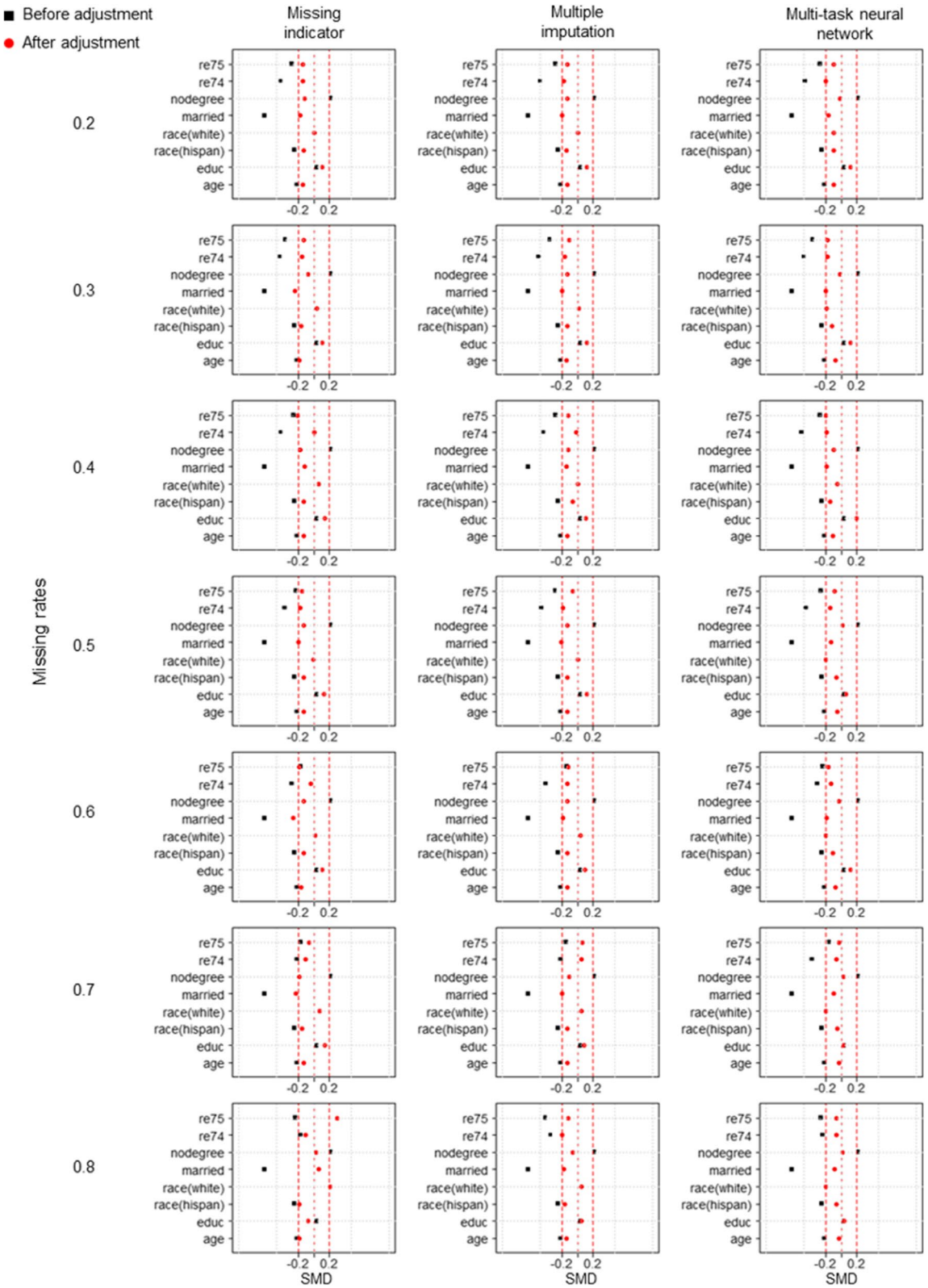

Fig. 6 Between-group standardized mean differences under MAR for covariates adjusted for propensity scores calculated by three different methods

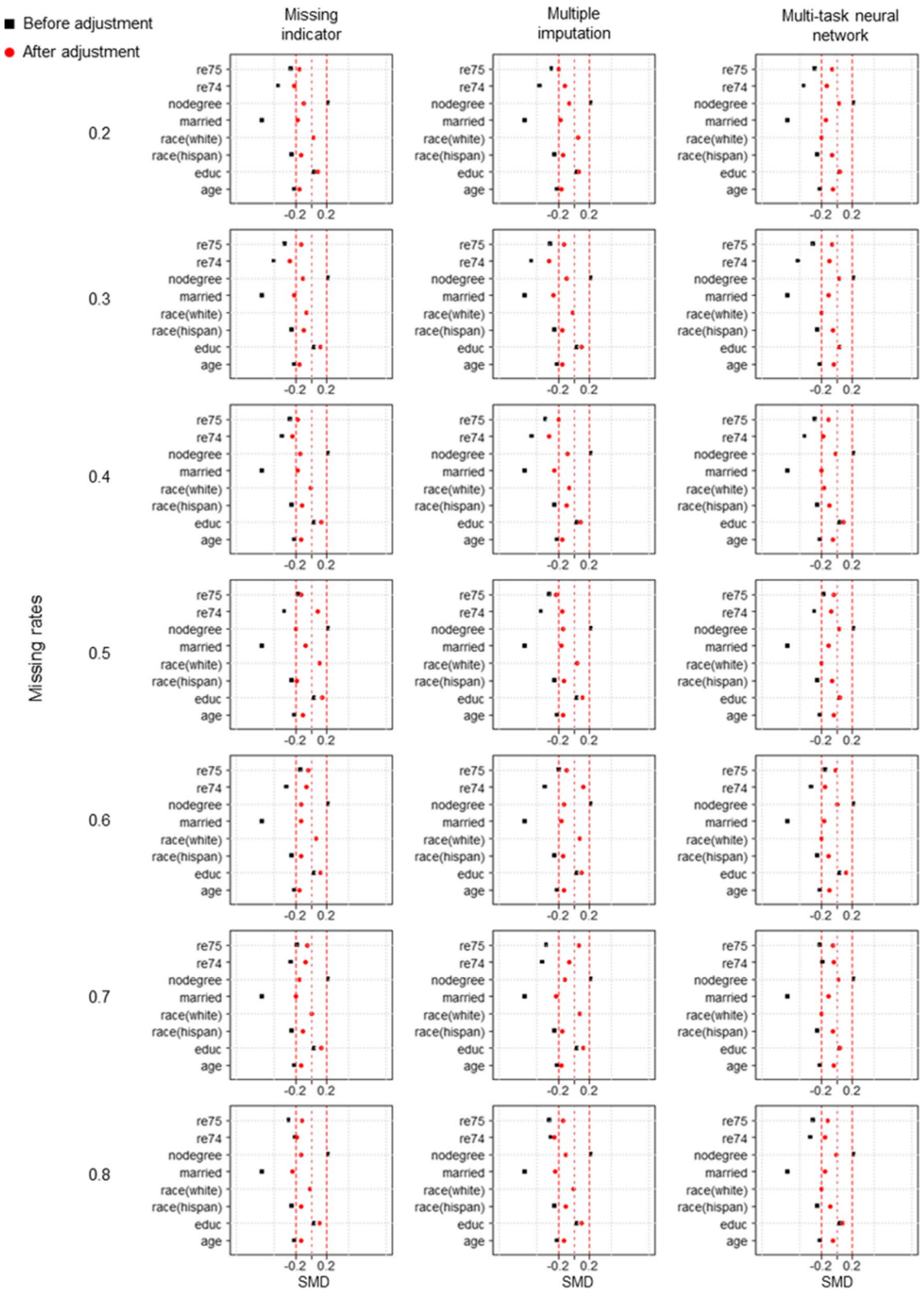

Fig. 7 Between-group standardized mean differences under MNAR for covariates adjusted for propensity scores calculated by three different methods

possible case, there is a slightly different result between the 2 types of data. Second, we cannot know the true effect of real-world data. Our model aims to establish a more accurate method for estimating model parameters when missing values are present. For this purpose, a complete real data modeling process is used as a standard of evaluation. It is our goal to prove that the proposed method can estimate the parameter value with the missing value as close as possible to the parameter value estimated without the missing value. Therefore, "true effect" should actually mean "effect estimated from full data" in real-world data. Third, MTNN assumes that input variables are correlated. Using the joint learning technique and the shared hidden layer, this correlation is used to estimate propensity scores and fill in missing values. When the input variables are independent or weakly correlated, MTNN may be unable to provide accurate estimates.

Conclusion

In this study, we propose a novel method for estimating propensity scores in data with missing values. It is based on a multi-task neural network, where missing value imputation and propensity score estimation are jointly trained as related tasks. Through the experimental results of simulated data and real-world data, we prove that our model has the smallest error in estimating the true effect under different missing mechanisms and different missing rates, and the standard deviation of the effect estimate is also the smallest. This shows that our method has good applicability in real-world observational studies with missing values.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-023-01847-2.

Additional file 1: Table S1 Variable descriptions for the real dataset. **Table S2** Summary of the real dataset. **Table S3** Estimation of the true effect in the simulated datasets using three different methods under the MAR mechanism. **Table S4** Estimation of the true effect in the simulated datasets using three different methods under the MNAR mechanism. **Table S5** Estimation of the true effect in the real datasets using three different methods under the MAR mechanism. **Table S6** Estimation of the true effect in the real datasets using three different methods under the MNAR mechanism. **Table S7** Regression coefficients for real-world data without missing values. **Table S8** Spearman's correlation coefficient for each input variable in real-world data.

Acknowledgments
The authors would like to thank Professor He Daihai for theoretical guidance.

Authors' contributions
Study conception and design: S Yang, J Luo and X Yan; Collection and creation of data: J Luo, P Du and S Yang; Data analysis and interpretation: S Yang, J Luo, X Yan, X Feng, P Du; Drafting the manuscript and figures: all authors; Final approval of manuscript: all authors.

Funding
This work was partially supported by the National Natural Science Foundation of China [grant number 11901352]; the Research Grants Council of the Hong Kong Special Administrative Region, China [HKU C7123-20G]; "Coronavirus Disease Special Project" of Xinglin Scholars of Chengdu University of Traditional Chinese Medicine [grant number XGZX2013].

Availability of data and materials
The data in this study is available from the corresponding author on reasonable request. Readers interested in the code of the simulation analysis may contact the corresponding author.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have no conflicts of interest to declare.

Received: 17 September 2022 Accepted: 20 January 2023
Published online: 15 February 2023

References

1. Webster-Clark M, Stürmer T, Wang T, Man K, Marinac-Dabic D, Rothman KJ, et al. Using propensity scores to estimate effects of treatment initiation decisions: state of the science. Stat Med. 2021;40(7):1718–35.
2. Austin PC, Jembere N, Chiu M. Propensity score matching and complex surveys [J]. Stat Methods Med Res. 2018;27(4):1240–57.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
4. Lin J, Gamalo-Siebers M, Tiwari R. Propensity-score-based priors for Bayesian augmented control design. Pharm Stat. 2019;18(2):223–38.
5. Cham H, West SG. Propensity score analysis with missing data. Psychol Methods. 2016;21(3):427.
6. D'Agostino RB Jr, Rubin DB. Estimating and using propensity scores with partially missing data. J Am Stat Assoc. 2000;95(451):749–59.
7. Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol. 2019;34(1):23–36.
8. Malla L, Perera-Salazar R, McFadden E, Ogero M, Stepniewska K, English M. Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review [J]. Journal of comparative effectiveness research. 2018;7(3):271–9.
9. Shao J, Wang L. Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika. 2016;103(1):175–87.
10. Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Stat Med. 2009;28(9):1402–14.
11. Crowe BJ, Lipkovich IA, Wang O. Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. Pharm Stat. 2010;9(4):269–79.
12. Mattei A. Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. Statistical Methods and Applications. 2009;18(2):257–73.
13. Linden A, Yarnold PR. Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. J Eval Clin Pract. 2016;22(6):875–85.
14. Cannas M, Arpino B. A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. Biom J. 2019;61(4):1049–72.
15. Tu C. Comparison of various machine learning algorithms for estimating generalized propensity score. J Stat Comput Simul. 2019;89(4):708–19.

16. Setoguchi S, Schneeweiss S, Brookhart MA, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study [J]. Pharmacoepidemiol Drug Saf. 2008;17(6):546–55.
17. Weberpals J, Becker T, Davies J, et al. Deep learning-based propensity scores for confounding control in comparative effectiveness research: a large-scale, real-world data study [J]. Epidemiology. 2021;32(3):378–88.
18. Kubat M. Neural networks: a comprehensive foundation by Simon Haykin Macmillan ISBN 0–02–352781-7. The Knowledge Engineering Review. 1999;13(4):409–12.
19. Caruana R. Multitask learning. Mach Learn. 1997;28(1):41–75.
20. Guo S, Fraser MW. Propensity score analysis: statistical methods and applications: SAGE publications; 2014.
21. Stuart EA. Matching methods for causal inference: a review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics. 2010;25(1):1.
22. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003;158(3):280–7.
23. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning [J]. Stat Med. 2010;29(3):337–46.
24. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. J Clin Epidemiol. 2010;63(8):826.
25. Santos MS, Pereira RC, Costa AF, et al. Generating synthetic missing data: a review by missing mechanism. IEEE Access. 2019;7:11651–67.
26. Garciarena U, Santana R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst Appl. 2017;89:52–65.
27. West SG, Cham H, Thoemmes F, et al. Propensity scores as a basis for equating groups: basic principles and application in clinical treatment outcome research. J Consult Clin Psychol. 2014;82(5):906.
28. Zhang P. Multiple imputation: theory and method. International Statistical Review/Revue Internationale de Statistique. 2003:581–92.
29. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. Jama. 2015;314(18):1966–7.
30. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46(3):399–424.
31. Lalonde RJ. Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev. 1986:604–20.
32. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J Am Stat Assoc. 1999;94(448):1053–62.
33. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? Epidemiology. 2018;29(2):191–8.
34. Wyss R, Schneeweiss S, Van Der Laan M, et al. Using super learner prediction modeling to improve high-dimensional propensity score estimation. Epidemiology. 2018;29(1):96–106.
35. Ju C, Combs M, Lendle SD, et al. Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. J Appl Stat. 2019;46(12):2216–36.
36. Choi BY, Wang C-P, Michalek J, et al. Power comparison for propensity score methods. Comput Stat. 2019;34(2):743–61.
37. Liu X. Methods and applications of longitudinal data analysis: Elsevier; 2015.

Publisher's Note