Computer Methods and Programs in Biomedicine 226 (2022) 107172

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Missing data imputation using utility-based regression and sampling approaches

Halimu N. Haliduola [a], Frank Bretz [b,c], Ulrich Mansmann [a,*]

[a] Institute for Medical Information Processing, Biometry and Epidemiology – IBE, LMU Munich, Munich, Germany
[b] Novartis Pharma AG, Basel, Switzerland
[c] Section for Medical Statistics, Medical University of Vienna, Vienna, Austria

ARTICLE INFO

Article history:
Received 1 October 2021
Revised 4 February 2022
Accepted 2 October 2022

Keywords:
Missing data
Utility-based regression
SMOTER
Machine learning

ABSTRACT

Data are often missing not at random (MNAR) in scientific experiments. We treat the MNAR problem as an imbalanced learning task. Standard predictive error measures of regression (e.g., mean squared error) are not suitable for imbalanced learning problems, such as in clinical trials where extreme values tend to be MNAR. We investigate hybrid imbalanced learning approaches that combine utility-based regression (UBR) with synthetic minority oversampling technique for regression (SMOTER) in cross-sectional trial settings. UBR optimizes the product of the conditional probability density (estimated by quantile regression forests) and a utility function which takes the relevance of the target variable value and the prediction error into account. SMOTER oversamples the relevant rare cases. Simulations show that the proposed method provides plausible predictions and reduces the bias for realistic missing data scenarios when compared with standard approaches like random forests and multiple imputation (systematic bias is observed in those methods, i.e., a tendency to underestimate the mean and standard deviation given the presence of MNAR in the area of high values of the target variable). The proposed method is implemented in a real dataset from an antidepressant clinical trial, and similar pattern of the systematic bias from commonly used methods is observed in the real data compare to the proposed method. Therefore, we encourage the integration of utility-based learning strategies for handling of missing data in the analysis of clinical trials.

1. Introduction

Missing data are the data that would be meaningful for the analysis but is not documented. The missing data, if not handled properly, will lead to lower statistical power for the analysis, and may lead to a bias in the estimated treatment effect and an underestimate of the variability. There are three types of missing mechanism [1]. (i) Missing Completely at Random (MCAR): if the probability of missingness does not depend on observed or unobserved measurements, e.g., patients move to another city due to non-health related reasons. (ii) Missing at Random (MAR): if the probability of missingness depends only on observed measurements conditional on the covariates in the model, e.g., younger people may more likely to have blood pressure not measured. (iii) Missing Not at Random (MNAR): if the probability of missingness depends on unobserved measurements, e.g., patients discontinue from the study due to lack of efficacy.

For the handling of missing data, many methods have been developed under the assumption of MAR or MNAR, respectively. In reality, however, missing data are often a mixture of different types. This makes the assumptions on the missing mechanism violated, which leads to poor performance of the handling methods [2]. To handle realistic missing data scenarios, Haliduola et al. [3] proposed a machine learning based missing data imputation framework where the MNAR problem is treated as an imbalanced learning task (since the MNAR cases are mostly distributed in one tail of the target variable). Take Fig. 1 as an example, depending on the proportion of MNAR data, regions that tend to have MNAR may have a smaller amount of available data than other regions (i.e., an imbalanced distribution). Imbalanced learning is necessary to compensate for the MNAR in that region and to avoid individual predictions being driven by the available non-missing data to an overall average level. They proposed oversampling of minority classes (i.e., the classes with extreme value of the target variable

* Corresponding author.
E-mail address: mansmann@ibe.med.uni-muenchen.de (U. Mansmann).

https://doi.org/10.1016/j.cmpb.2022.107172

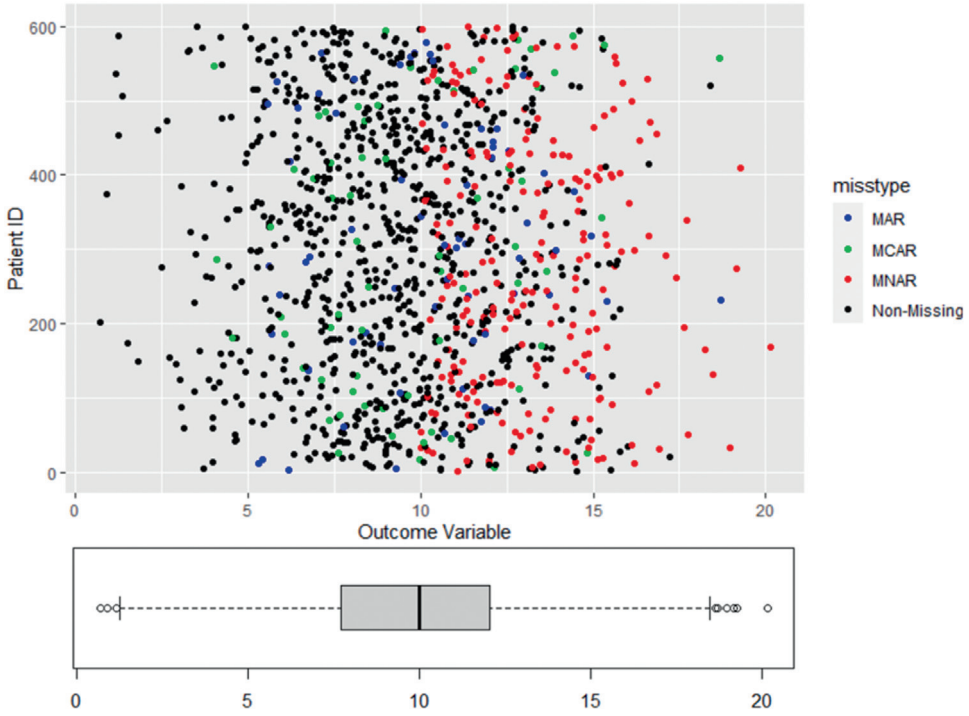

Fig. 1. Simulation data: scatter plot and boxplot for the target variable. The back dots are the non-missing data, blue dots are the MAR, green dots are MCAR, Red dots are MNAR. The details of the data generation process are described in Section 3.1.

that tend to be MNAR), followed by recurrent neural networks to model the data. This framework was shown to be effective for the handling of the missing data based on simulation studies and a real clinical trial data.

Haliduola et al. [3] used a simple random oversampling with replacement and a standard error measure (i.e., mean squared error, MSE). However, these methods come with drawbacks. First, in a simple random oversampling with replacement, random sets of copies of minority class cases are added to the data, which may lead to many duplicates in the minority class. During the learning process, the decision region for the minority class may become very specific and the model will give more focus in that region. For example, in a tree-based learning process, this may lead to new splits in the decision trees, which will result in more terminal nodes (leaves) as the learning algorithm tries to learn more and more specific regions of the minority class, and eventually this will cause overfitting of the model [5]. Secondly, the standard predictive error measure like MSE is not suitable for a regression problem with imbalanced distribution of target variable values in the training data (like in MNAR problem where the extreme values tend to be missing). Their weakness is that they are not sensitive to the location of target variable values [4]. See Fig. 1 as an example, considering the MNAR data (red dots), distribution of available non-missing data (black dots) are imbalanced across the range of target variable (i.e., less available data in the area of high values due to MNAR). If the error measure is not sensitive to the location of target variable values, the area of high values will get less focus in the training process due to the smaller amount of data in that area, and thus the impact of missing data on the aggregated estimation will be ignored. In such cases, it is important to give more focus on the area of high values in the training process to compensate for the MNAR and to avoid the prediction being driven by the frequent cases in the other locations of the target variable. Therefore, it is necessary to have an error metric that is sensitive to the location of the errors, which copes with imbalanced distribution of target variable values.

In this paper, to avoid model overfitting caused by the simple random oversampling, we use the synthetic minority oversampling technique for regression (SMOTER) [7] to oversample the relevant rare cases; and, to overcome the drawbacks of standard error measure, we use the imbalanced learning technique utility-based regression (UBR) [6], which takes both relevance (or importance) of the target variable values and the prediction errors into account in the optimization process. For simplicity, we consider cross-sectional data only. Quantile regression forests [9] are used to estimate the conditional probability density. The optimization process involves determining the maximum integral of the product of the conditional probability density function and the utility function for each case. In light of the "evidence-based computational statistics" [11,12], we evaluate the proposed method in an extensive simulation study using realistic missing data scenarios (i.e., mixture of MCAR, MAR, and MNAR data). The performance of proposed method is evaluated comprehensively in terms of the central tendency and variability of imputed data, prediction accuracy, and a performance comparison with commonly used methods like random forests and multiple imputation. Finally, we illustrate the proposed method with a real dataset from an antidepressant clinical trial, which is one of the few publicly available datasets that can be used to demonstrate methods for handling missing data where a continuous outcome is measured.

2. Methods

In this paper, we aim to handle realistic missing data scenario (i.e., mixture of MCAR, MAR, and MNAR data) in a continuous outcome variable. We treat the MNAR problem in clinical trials as an imbalanced learning task. We investigate a hybrid imbalanced learning approach that combines utility-based regression (UBR) [6] with synthetic minority oversampling technique for regression (SMOTER) [7] in the missing data imputation. First, we assign a relevance to the target variable values based on their distribution in the training data (i.e., available non-missing data) and

define a threshold for oversampling. The second step is data preprocessing, where we use the SMOTER method to oversample cases with relevance greater than the threshold. In the third step, we apply utility-based regression on the oversampled training data, and the model parameters are optimized by maximizing the relevance and minimizing the error simultaneously. The final step is to use the optimal model to predict the missing target variable values.

2.1. Utility-based regression

Let Y be a target variable and X predictor vector. The most commonly used error measures in regression are the mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)^2,$$

and the mean absolute error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\bar{y}_i - y_i|.$$

The standard predictive error measures are not suitable for a regression problem with imbalanced distribution of target variable values in the training data (like in MNAR problem where the extreme values tend to be missing). Their weakness is that they are not sensitive to the location of target variable values [4]. Take Fig. 1 as an example, if the error measure is not sensitive to the location of target variable values, the area of high values will get less focus in the training process due to the smaller amount of data in that area (due to MNAR), and thus the impact of missing data on the aggregated estimation will be ignored. In such cases, it is important to give more focus on the area of high values in the training process to compensate for the MNAR and to avoid the prediction being driven by the frequent cases in the other locations of the target variable. Therefore, it is necessary to have an error metric that is sensitive to the location of the errors, which copes with imbalanced distribution of target variable values. It should be noted that the example in Fig. 1 is used to demonstrate the idea, it could be the other way round in reality, i.e., the lower values tend to be MNAR.

Utility is a function of both the error of the prediction and the relevance (or importance) of both true and predicted values. Together, the relevance and loss information give a utility function, which provides more reliable evaluation of a regression model. The ultimate goal of utility-based regression is to maximize the utility, which is achieved by maximizing the relevance and minimizing the error simultaneously. In following sections, we use the notations for utility-based regression defined by Torgo and Ribeiro [6] and Ribeiro [4].

The relevance is the crucial property that distinguishes non-uniform cost/benefit regression problems from those standard regression problems. The relevance function $\emptyset(Y)$: $y \rightarrow [0,\ 1]$ is a continuous function that expresses the domain-specific importance concerning the target variable domain y by mapping it into a $[0,\ 1]$ scale of relevance, where 0 represents the minimum and 1 represents the maximum (see an example in Fig. 2). To take both predicted value ($\bar{y}$) and true value (y) into account, the joint relevance function is defined as weighted average:

$$\emptyset(\bar{y},\ y) = (1-p)\ \emptyset(\bar{y}) + p\ \emptyset(y)$$

where $p \rightarrow [0,\ 1]$ is the weight, e.g., $p = 0.5$. The actual form the relevance function is domain specific and defined by the user based on the problem in hand.

For the missing data problem in a continuous target variable (like the example mentioned above), a relevance function can be defined to assign more relevance/importance to the extreme values in one tail or both tails according to the distribution of available data. For example, we use boxplot whiskers or summary statistics like the first quartile (Q1) and the third quartile (Q3) to identify the extreme values. In the example in Fig. 2, the extreme values are identified using the boxplot whiskers, and then the maximum relevance of 1 is assigned to the extreme cases, and minimum relevance of 0 is assigned to the median value. A monotone cubic spline interpolation line over a set of maximum and minimum relevance points is the actual shape of the relevance function [8]. Using the boxplot to identify the extreme values, a coefficient needs to be specified to determine how far the whiskers extend to the extreme data points in the boxplot (e.g., a coefficient of 1.5 as in the standard boxplot). The choice of the coefficient should be based on the specific problem in hand and it should be prespecified. For example, a coefficient smaller than 1.5 can be considered to assign high relevance to more data points. A range of the coefficients can also be considered to perform the sensitivity analysis. In our example, considering the presence of MNAR in the area of high values and the MCAR/MAR spread out across the whole range of target variable, it makes sense to assign more relevance for both high and low extreme values. Assigning high relevance in both tails may also avoid disproportionately heavy in one tail over the other in the prediction.

The cost of a prediction is defined as product of the relevance and the loss (or error) function,

$$c(\bar{y},\ y)\ =\ \emptyset(\bar{y},\ y)\ C_{max}\ L(\bar{y},\ y)$$

where $\emptyset(\bar{y},\ y)$ is the joint relevance function, C_{max} is the maximum cost that is only assigned when the relevance is maximum (i.e., $\emptyset(\bar{y},\ y) = 1$). The term $\emptyset(\bar{y},\ y)\ C_{max}$ can be seen as a case-specific maximum cost value, i.e., the maximum penalty we get if $\bar{y}$ is the "worst possible" prediction for the particular case under consideration. $L(\bar{y},\ y)$ is the loss function. It is important to scale the loss function to [0, 1]. Torgo and Ribeiro [6] defined a percentage-type loss function as the difference between the maximum and minimum relevance in the interval between the true and predicted values.

$$L(\bar{y},\ y) = [\max_{i\in\bar{y}..\ y}\emptyset(i) - \min_{i\in\bar{y}..\ y}\emptyset(i)]$$

The total cost can be calculated by summing up all individual cost values. It is important to notice that when asserting the cost of a prediction, it is necessary to take both the true and the predicted values into account. Predicting an irrelevant value for a case that has an actual extreme value is not the only cost that can occur. It may be equally serious to predict an extreme value for a frequent case, as it causes false alarm that could lead to serious cost. Therefore, the joint relevance function is used in the above cost function. In addition, it makes sense to use weight $p = 0.5$ in the joint relevance function to give equal importance to both types of error.

The benefit of a prediction is defined as product of the relevance of true value and the complementary of the loss,

$$b(\bar{y},\ y) = \emptyset(y)\ B_{max}\ (1 - L(\bar{y},\ y))$$

where $\emptyset(y)$ is the relevance function of true value, B_{max} is the maximum reward that is only assigned when the relevance is maximum. In the benefit function, only the relevance of the true value is considered as the purpose is to assert how well a model predicts the test cases that are relevant (i.e., rewards the accurate prediction for the relevant values). The total benefit can be calculated by summing up all individual benefit values.

The utility of a prediction is the net balance between its benefits and costs, defined as,

$$U(\bar{y},\ y)\ = b(\bar{y},\ y)\ - c(\bar{y},\ y)$$

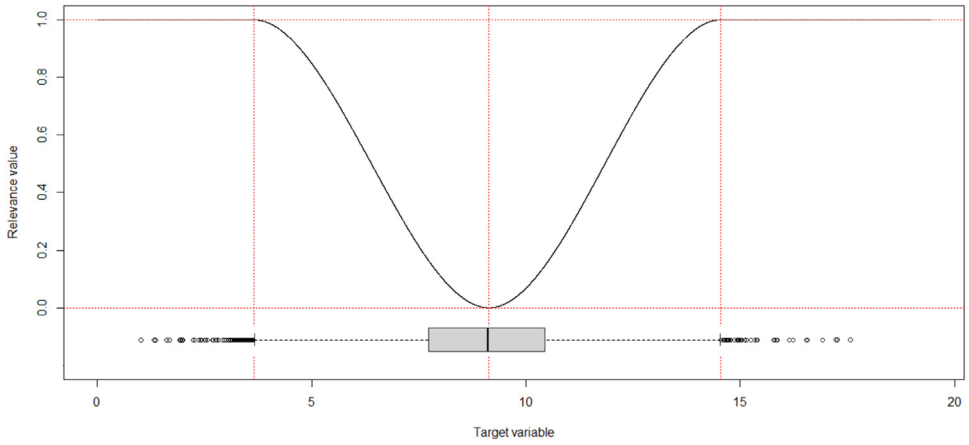

Fig. 2. An example of relevance function to assign more importance to the extreme values according to the distribution of available data.

The total utility can be calculated by summing up all individual utility values. The mean utility can also be calculated as utility-based model performance metrics.

2.2. Quantile regression forests

As mentioned in Section 2.1, the ultimate goal of utility-based regression is to optimize the utility, which is achieved by maximizing the relevance and minimizing the error simultaneously. In this paper, we use the optimization process proposed by Rau et al. [14]. This method uses quantile regression forests (QRF, [9]) to estimate the conditional probability density which is a crucial element in the optimization process. To elaborate the main idea of QRF, we start with the random forests (RF, [16]) and quantile regression [18].

The random forests build k trees in parallel using n independent observations (y_i, x_i), $i = 1, \ldots, n$. Each tree is based on the bootstrapped data (random sampling with replacement, e.g., use 2/3 as the original data size) and random subset of variables (e.g., use 1/3 of all feature variables). This kind of variety is what makes random forests more effective than individual decision tree. Let θ denote the random parameter vector that determines how a tree is grown (e.g., which variables are considered for split points at each node), the corresponding tree is denoted by $T(\theta)$, let Lf denote the leaves of the tree ($L = 1, \ldots, m$). For every $x \in X$, there is only one leaf Lf can be obtained when dropping x down the tree. Denote this leaf by $Lf(x, \theta)$ for tree $T(\theta)$. For a single tree, the weight vector $w_i(x, \theta)$ is a positive constant if observation x_i is part of leaf $Lf(x, \theta)$ and 0 if not, and the weights $w_i(x, \theta)$ sum to 1. The prediction of a single tree k, given the feature $X = x$, is then weighted average of the original observations y_i,

$$\bar{u}_t(x) = \sum_{i=1}^{n} \omega_i(x, \theta) y_i,$$

where t is the tth single tree, $t = 1, \ldots, k$. The conditional mean $E(Y|X = x)$ is approximated by the averaged prediction of k single trees, each constructed with an independent and identically distributed vector θ_t. Let $\omega_i(x)$ be the average of $\omega_i(\theta)$ over the trees, defined as,

$$\omega_i(x) = k^{-1} \sum_{t=1}^{k} \omega_i(x, \theta_t).$$

The predictions of random forests are then the weighted conditional mean,

$$\bar{u}(x) = \sum_{i=1}^{n} \omega_i(x) y_i.$$

The weighted conditional mean is estimated by minimizing the MSE:

$$E(Y|X = x) = arg \min_{\bar{y}} E\{(\bar{y} - y)^2 | X = x\}.$$

The conditional mean describes only one aspect of the conditional distribution of a target variable Y, while the quantile regression aims to provide more information about the conditional distribution, e.g., the conditional quantiles [18]. For $X = x$, the conditional distribution function $F(y|X = x)$ is given by the probability of Y is smaller than $y \in \mathbb{R}$ ($\mathbb{R}$ is the space for the target variable),

$$F(y|X = x) = P(Y \leq y|X = x).$$

For a continuous distribution function, given $X = x$, the α-quantile $Q_\alpha(x)$ is then defined such that the probability of Y being smaller than $Q_\alpha(x)$ is exactly equal to α $(0 < \alpha < 1)$. The quantiles $Q_\alpha(x)$ give more information about the conditional distribution of Y, which is defined as,

$$Q_\alpha(x) = \inf\{y : F(y|X = x) \geq \alpha\}.$$

The loss function L_α is defined as the weighted absolute deviations,

$$L_\alpha(y, q) = \begin{cases} \alpha|y - q| & y \rangle q \\ (1 - \alpha)|y - q| & y \leq q \end{cases}$$

The conditional quantiles are estimated by minimizing the expected loss $E(L_\alpha)$,

$$Q_\alpha(x) = arg \min_{q} E\{L_\alpha(Y, q)|X = x\}.$$

For quantile regression forests, trees are grown as in the standard random forests algorithm [9]. The conditional distribution is then estimated by the weighted distribution of observed target variables, where the weights $(\omega_i(x))$ attached to observations are identical to the original random forests algorithm. The key difference from the standard random forests is that, for each node in each tree, QRF keeps the value of all observations in this node (not just their mean as in the standard random forests), and assesses the conditional distribution of those observations.

For $X = x$, the conditional distribution function of Y is given by,

$$F(y|X = x) = P(Y \leq y|X = x) = E(I_{\{Y \leq y\}}|X = x)$$

where $I_{\{Y \le y\}}$ is the indicator function, which equals to 1 if $Y \le y$ otherwise 0. Just as $E(Y|X = x)$ is approximated by a weighted mean of Y, define an approximation to $E(1_{\{Y \le y\}}|X = x)$ by the weighted mean over the observations of $1_{\{Y \le y\}}$ as the prediction of QRF,

$$\bar{F}(y|X = x) = \sum_{i=1}^{n} \omega_i(x) 1_{\{Y \le y\}}.$$

The optimization process uses a method proposed by Rau et al. [14], which use QRF to estimate the conditional probability density. In regression, for each case, this process involves determining the maximum integral of the product of the conditional probability density function and the utility function. The optimal prediction for $X = x$ is given by,

$$\bar{y}(X = x) = arg\ max[\bar{y}] \int pdf(y|X = x)\, U(\bar{y},\ y)\, dy$$

where $pdf(y|X = x)$ is the conditional probability density estimation for $X = x$, and $U(\bar{y},\ y)$ is the utility evaluated on the true value y and predicted value $\bar{y}$. Final predictions are the conditional means take target variable utility into account. We use the R package "UBL" (stands for "Utility-Based Learning", [13,15]) in this paper.

2.3. SMOTER

Synthetic Minority Oversampling Technique (SMOTE) was introduced by Chawla et al. [5] for the classification task. This algorithm operates in the feature space rather than target variable space (as all rare cases have the same target minority class). The minority class is oversampled by taking each minority sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors (e.g., $k = 5$). For example, if the amount of oversampling needed is 200%, only two neighbors from the k nearest neighbors are chosen and one sample is generated in the direction of each. Synthetic samples are generated in the following way: take the difference between the feature vector under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and then add it to the feature vector under consideration.

Torgo et al. extended the SMOTE for regression task (i.e., the SMOTER) in 2013. Three key components were addressed in the extension: the relevance function (i.e., the $\phi(Y)$ as discussed in Section 2.1) and the user-specified threshold for the relevance values were used to define the relevant (rare) cases and the frequent cases (e.g., relevance threshold = 0.8); the same approach as in the original algorithm was used to generate the synthetic feature samples; the weighted average of the target variable values of the two seed examples (i.e., the case and the selected neighbor) was used as the synthetic value for the target variable (the weights are calculated as an inverse function of the distance of the generated new case to each of the two seed examples). We use the R package "UBL" [15] for the implementation of SMOTER.

In practice, it is common to implement the SMOTER together with the undersampling of frequent cases. However, in this paper, we do not consider the undersampling for following reasons: 1) In realistic missing data scenarios, the MNAR data are located in certain area of the target variable, but the MAR/MCAR data may spread out in the whole range of the target variable. In the training process, it may not be a conservative approach to give less focus on the locations where MAR/MCAR data may appear; 2) The undersampling reduces the size of the training data, this may not be a favorable approach in clinical trials in which the total amount of data is normally not massive.

In SMOTE, the amount of oversampling is a hyperparameter of the system [5]. We fine-tuned the appropriate amount of oversampling using the cross-validation (CV) approach. It is important to note that only the training data should be oversampled during the CV process, the validation data should never be oversampled to avoid the "overoptimism" issue [19].

	MCAR	MAR	MNAR	OUT-COM	COV1	COV2	COV3	COV4	COV5	COV6	COV7
MCAR	1	0	0	0	0	0	0	0	0	0	0
MAR	0	1	0	0	0.4	0	0	0	0	0	0
MNAR	0	0	1	0.5	0	0.2	0.2	0.2	0.2	0.2	0.2
OUTCOME	0	0	0.5	1	0	0.5	0.5	0.5	0.5	0.5	0.5
COV1	0	0.4	0	0	1	0	0	0	0	0	0
COV2	0	0	0.2	0.5	0	1	0.2	0.2	0.2	0.2	0.2
COV3	0	0	0.2	0.5	0	0.2	1	0.2	0.2	0.2	0.2
COV4	0	0	0.2	0.5	0	0.2	0.2	1	0.2	0.2	0.2
COV5	0	0	0.2	0.5	0	0.2	0.2	0.2	1	0.2	0.2
COV6	0	0	0.2	0.5	0	0.2	0.2	0.2	0.2	1	0.2
COV7	0	0	0.2	0.5	0	0.2	0.2	0.2	0.2	0.2	1

Fig. 3. Example of correlation matrix used in the simulation data generation.

3. Simulation study to evaluate performance of methods

3.1. Design of simulation study

To demonstrate the idea of the utility-based regression and sampling approaches, we consider the cross-sectional data only in this paper. In the simulation study, random data is generated for 600 subjects. The outcome variable and covariates (predictors) are normally distributed. Missing data indicators are binary variables (i.e., separate indicator variables for MCAR, MAR and MNAR). Correlated normal and binary data are generated simultaneously using the point-biserial correlation approach of Demirtas and Dogana [20]. Suppose that X and Y follow a bivariate normal distribution with a correlation of ρ_{XY}. If X is dichotomized to produce X_D, then the resulting correlation between X_D and Y can be given as point-biserial correlation,

$$\delta_{X_D Y} = \rho_{XY} \left(\frac{h}{\sqrt{p(1-p)}} \right)$$

where p is the proportion of the observations above the point of dichotomization, and h is the ordinate (probability density function) of the normal curve at the same point.

In the simulation study, we simultaneously generate one outcome variable and seven covariates (each normally distributed with mean 10 and variance 10) and 3 missingness indicators using a given correlation matrix (see Fig. 3 for an example). The MCAR flag (with missing data proportion of 5%) is independent from any other variables. The MAR flag (with missing data proportion of 5%) is correlated with the first covariate only (correlation coefficient = 0.4) and independent from the outcome variable and the other covariates. To evaluate the performance of imputation method properly, we consider higher proportion of MNAR data (i.e., 25%). The MNAR flag is positively correlated with outcome variable (i.e., the higher values tend to be missing, correlation coefficient = 0.5) and the second to seventh covariates (correlation coefficient = 0.2). The outcome variable is correlated with the MNAR flag and the second to seventh covariates (correlation coefficient = 0.5). The first covariate is correlated with MAR flag only. The second to seventh covariate are correlated with the outcome, therefore they are also correlated with each other (correlation coefficient = 0.2). See Fig. 1 as an example for the distribution of the outcome variable. We use the R package "BinNor" [21] in the data generation. Since the higher values of outcome variable tend to be missing (MNAR), the mean of the available non-missing data is an underestimation of the true value. A proper missing imputation method should compensate for the MNAR and reduce the bias

in the aggregated estimation. In this paper, we perform the simulation with 100 replications.

We impute the missing data using proposed method, i.e., UBR facilitated by SMOTER (ubr.smt). In the SMOTER process, we identify the relevant extreme values based on the summary statistics of available training data, i.e., the data points $\leq$ the first quartile (Q1) or $\geq$ the third quartile (Q3) are oversampled. The amount of oversampling is determined as 3 times as the available data in both tails based on the cross-validation. In the UBR process, we assign relevance function to target variable using the boxplot with a coefficient of 0.75 (i.e., half of the standard coefficient). Based on the summary statistics of available data, a coefficient of 0.75 is considered as appropriate to assign relevance to the high target variable values where tend to have MNAR and also the low extreme values. A range of coefficients (0.5, 0.6, 0.7, 0.8 and 0.9) are also experimented to illustrate the impact of relevance function on the imputation performance. As mentioned above, the relevance function is defined according to the distribution of the available data, and there is a shift in the central tendency of the available data due to MNAR in the area of high values. This shift is also reflected in the relevance function, which leads to more relevance given in the area of high values (this is considered as a conservative approach given the presence of MNAR in that area only in this case).

3.2. Measuring performance of the proposed methods

To compare the performance of proposed method (i.e., UBR facilitated by SMOTER), we impute the missing data using other methods including:

- ubr.org = UBR without facilitating by SMOTER.
- qrf.smt = QRF facilitated by SMOTER, details of QRF are described in Section 2.2. We use the R package “quantregForest” [10] in the implementation.
- qrf.org = QRF without facilitating by SMOTER.
- rf.smt = random forests facilitated by SMOTER, details of RF are described in Section 2.2. We use the R package “randomForest” [17] in the implementation.
- rf.org = random forests without facilitating by SMOTER.
- mi = traditional multiple imputation under the assumption of MAR. In addition to those machine learning-based methods, comparisons with the most commonly used traditional statistical methods (i.e., multiple imputation) are also considered meaningful. We use the R package “MICE” (van Buuren et al. [24]) with 200 multiple imputations. MICE stands for Multivariate Imputations by Chained Equations, which generates multiple imputations for incomplete multivariate data by Gibbs sampling. The algorithm imputes an incomplete target column by generating "plausible" synthetic values given other columns (covariates) in the data. The imputation method for the missing continuous outcome variable is predictive mean matching ([22] and [23]).

We perform the following measures to compare the performance of difference methods:

- Calculate the mean and standard deviation (SD) of the imputed outcome variable by different imputation methods as mentioned above, and compare with the mean and SD of true value (i.e., the complete outcome variable before set the missing values). If the estimations are close to the mean and SD of true value then the imputation method is appropriate. To show the bias that caused by missing data, the mean and SD of available non-missing data are also provided.
- Perform one sample *t*-test on the imputed data with a null-hypothesis of mean = 10, the larger p-values indicate better imputation performance.
- Perform a simple linear regression of imputed value versus the true value, and compare the intercepts (close to 0 is better) and the slops (close to 1 is better).

3.3. Simulation results

We visualize the performance measures from 100 studies using the boxplot. In Fig. 4, the boxplots for the mean values from 100 studies per scenario are presented. The true means follow normal distribution around 10 (the blue box). The bias caused by the missing data is substantial, the means estimated from non-missing available data are significantly lower than the true means (i.e., noimp, the brown box on the right in below figure). The means estimated based on imputed data by the proposed method (i.e., UBR + SMOTER) are the closest to the true means (the green box) when comparing with other methods. The means from the UBR without SMOTER (the light green box) are the second closest estimation of the true means. The QRF and RF perform very similarly (the boxes labeled as qrf.org and rf.org), which is expected as the goal is to provide the conditional mean as prediction. When facilitating by SMOTER, QRF and RF perform better than without SMOTER but still are not as good as the proposed method (the boxes labeled as qrf.smt and rf.smt). The traditional multiple imputation is not as good as the proposed method (the purple box). In general, all other methods tend to underestimate the mean given the presence of MNAR in the area of high values of the target variable.

It is also important to evaluate the performance of imputation method in terms of the variability of imputed data. As shown in Fig. 5, similar as for the central tendency measure (i.e., the mean), the proposed method provides the closest estimation for the SD, followed by the UBR without facilitating by SMOTER. All other methods tend to underestimate the SD given the presence of MNAR in the area of high values of the target variable.

We perform sensitivity analysis in terms of the coefficient of the relevance function. A range of coefficients (i.e., 0.5, 0.6, 0.7, 0.8 and 0.9) are experimented and results are shown in Fig. 6 (A for distribution of mean and B for distribution of SD). It is clear that the relevance function impacts the performance of UBR considerably. The coefficient here is a parameter to determine how far the whiskers extend to the extreme data points in the boxplot when defining the relevance function. The higher coefficients result in high relevance been assigned to the more extreme cases (e.g., for less data points), this may increase the variability of the predicted values. As mentioned in Section 3.1, there is a shift in the relevance function due to MNAR in the area of high values, this leads to even less lower extreme values been assigned with high relevance. Therefore, higher coefficients result in higher estimated mean and SD in this case. As mentioned above, all commonly used methods tend to underestimate the mean and SD given the presence of MNAR in the area of high values of the target variable. It would be equally worse to overestimate the mean and SD (e.g., in the case of coefficient = 0.9). Therefore, it is important to pre-specify a proper relevance function according to the distribution of available data and make a plausible assumption on the missing data (i.e., the possible locations of target variable scale where the missing data tend to occur). It is also important to perform sensitivity analysis with different relevance functions (and associated parameters) to check the appropriateness and robustness of the primary analysis.

We perform one sample *t*-test on the imputed data (imputed by different methods) with a null-hypothesis of mean = 10 and present the distribution of p-values in Fig. 7. For the true data (where no missing data), the p-values are mostly greater than 0.05 as expected. For the proposed method (i.e., UBR + SMOTER), the majority of the p-values are greater than 0.05. While for other

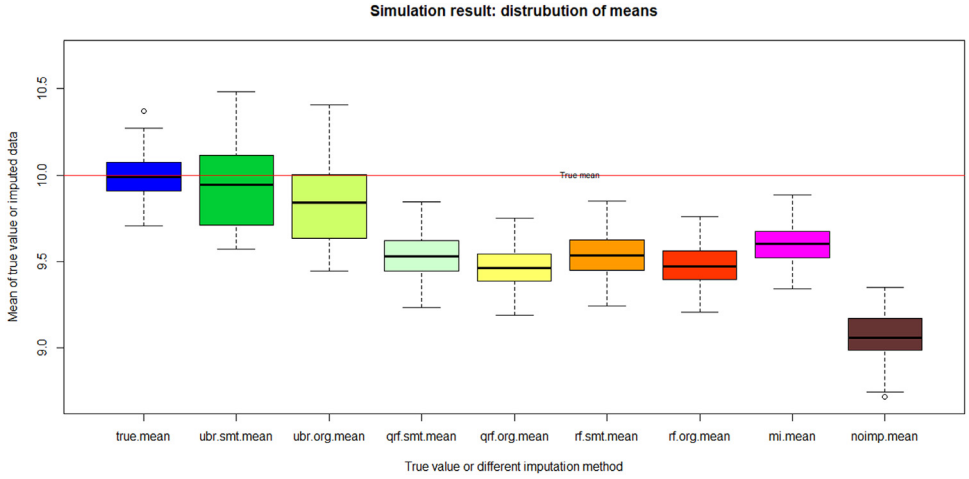

Fig. 4. Simulation result – distribution of means of imputed data by different methods. ubr = utility-based regression (coefficient of relevance function = 0.75), smt = SMOTER data, org = original data, qrf = quantile random forests, rf = random forests, mi = multiple imputation, noimp = no imputation.

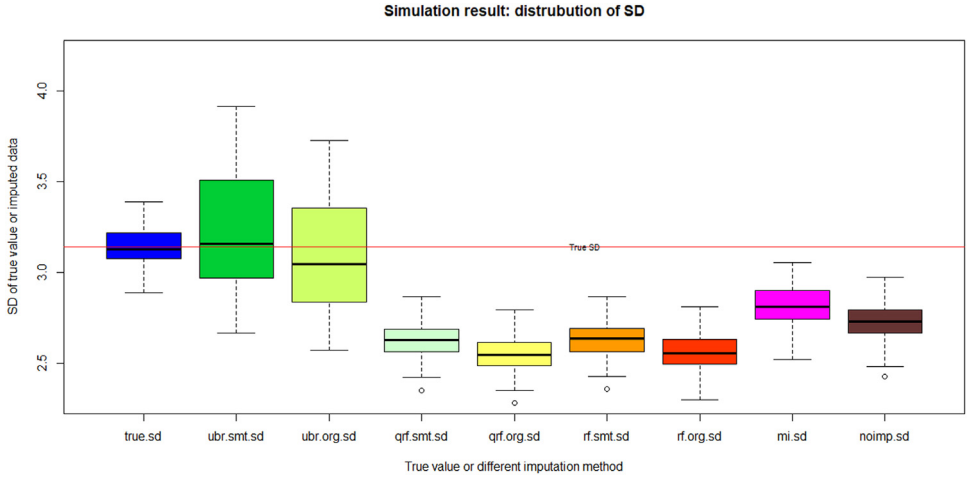

Fig. 5. Simulation result – distribution of SDs of imputed data by different methods. ubr = utility-based regression (coefficient of relevance function = 0.75), smt = SMOTER data, org = original data, qrf = quantile random forests, rf = random forests, mi = multiple imputation, noimp = no imputation.

methods, the p-values are quite small (mostly < 0.05). Although the p-value is sample size dependent, but the trend is clear to show that the proposed method is better than other methods in terms of the ability to reduce the bias of missing data in the aggregated estimation.

We perform simple linear regression for the true value versus the imputed value (by different method). The intercept and the slop from the linear regression are visualized using the boxplots in Fig. 8. The proposed method (i.e., UBR + SMOTER) gives the least intercept and the greatest slop (i.e., closest to 1), suggesting the best performance among all the methods.

4. Real data example

We implement the proposed method in a real dataset from an antidepressant clinical trial, which is available on the website of London School of Hygiene and Tropical Medicine [25]. Original data are from an antidepressant clinical trial with four treatments; two doses of an experimental medication, a positive control, and placebo [26]. There are 26.1% and 25.0% patients with missing Hamilton 17-item rating scale for depression (HAMD17) at Week 6 in Control group (i.e., placebo, $N = 88$) and Test group (i.e., created by randomly selecting patients from the three non-placebo arms, $N = 84$), respectively.

We use the HAMD17 at Week 6 as the target variable (cross-sectional data), use the treatment group and the available baseline variables as predictors (including the gender, baseline HAMD17 value, HAMD Total score and Patient Global Impression of Improvement (PGI-I)). The reasons for discontinuation are not available in the published dataset, this makes it difficult to make assumption about the missing mechanism. We define the relevance function according to the summary statistics of the available data. In the pre-processing, the data points $\leq$ Q1 or $\geq$ Q3 are oversampled using SMOTER method. The amount of oversampling is determined as twice as the original available data in both tails based on the cross-validation. In the UBR process, maximum relevance of 1 is assigned to the data points $\leq$ Q1 or $\geq$ Q3, and minimum relevance of 0 is assigned to median value (note: it is not the boxplot method in this case and therefore no coefficient to be determined). A monotone cubic spline interpolation line over a set of maximum and minimum relevance points is the actual shape of the relevance function.

To compare the imputation performance, we impute the missing data using the methods as described in Section 3.2. The imputed outcome variable (i.e., change from baseline in HAMD17 score at Week 6) is analyzed using the analysis of covariance (ANCOVA) model with treatment as factor and baseline value as covariate. To show the bias that caused by the missing data, we also

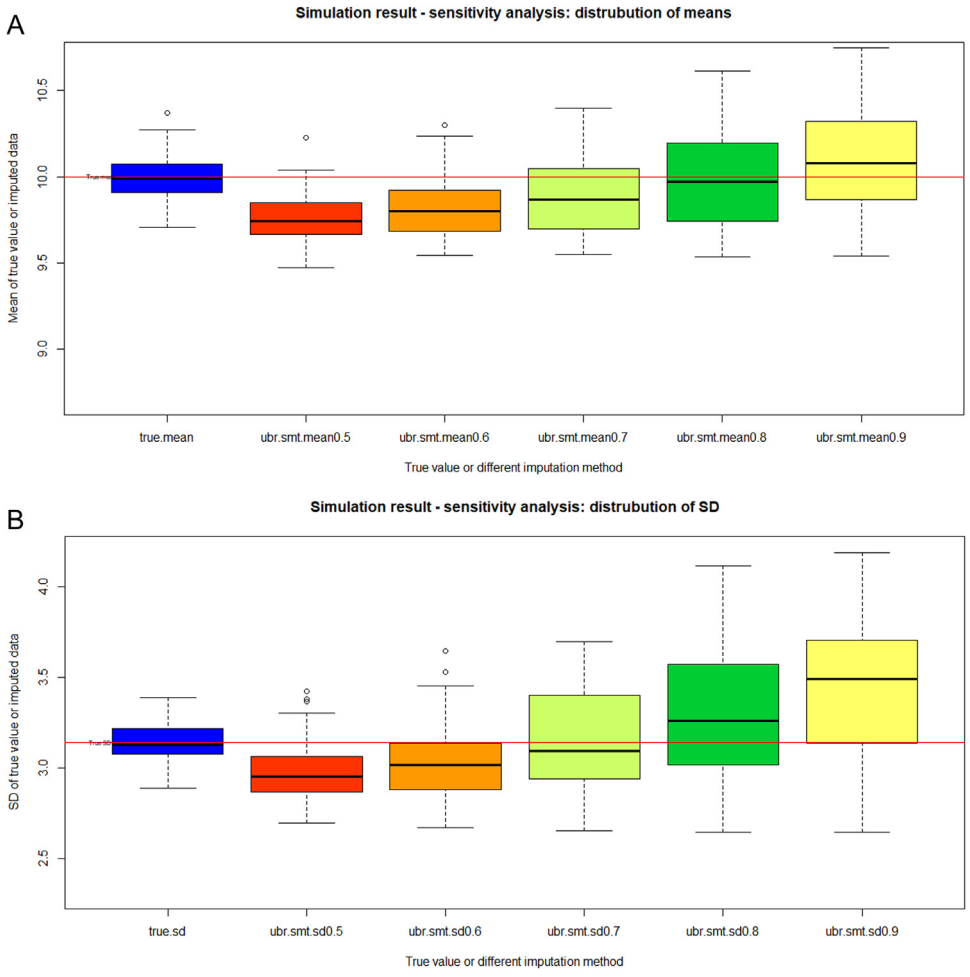

Fig. 6. Simulation result – sensitivity analysis: distribution of means (A) and SDs (B) of imputed data by ubr+smt using different coefficients in the relevance function (0.5, 0.6, 0.7, 0.8, 0.9). ubr = utility-based regression, smt = SMOTER data.

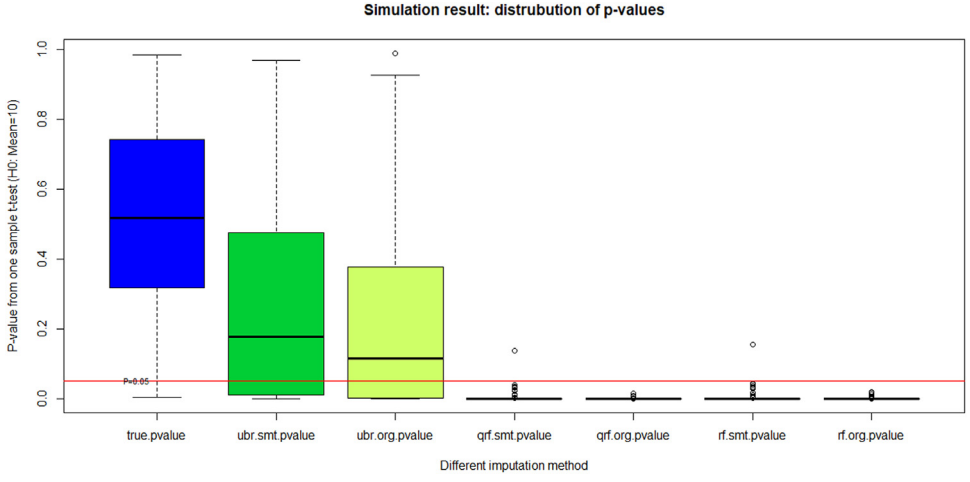

Fig. 7. Simulation result – distribution of p-values from one sample *t*-test on imputed data by different methods. ubr = utility-based regression (coefficient of relevance function = 0.75), smt = SMOTER data, org = original data, qrf = quantile random forests, rf = random forests.

analyze the outcome variable without imputation using ANCOVA. The results from the different approach are presented in forest plot (Fig. 9). The proposed imputation method (i.e., UBR + SMOTER) provided the most conservative estimation for the treatment effect in both treatment groups. There is systematic bias in the results from other methods. This bias is more pronounced in the Control group, a possible reason could be there are more low responders with missing data in Control group (e.g., may be more MNAR in Control group). In general, comparing with the proposed method, other methods tend to be optimistic, which may lead to aggressive estimation and hence introduce bias in the study conclusion (es-

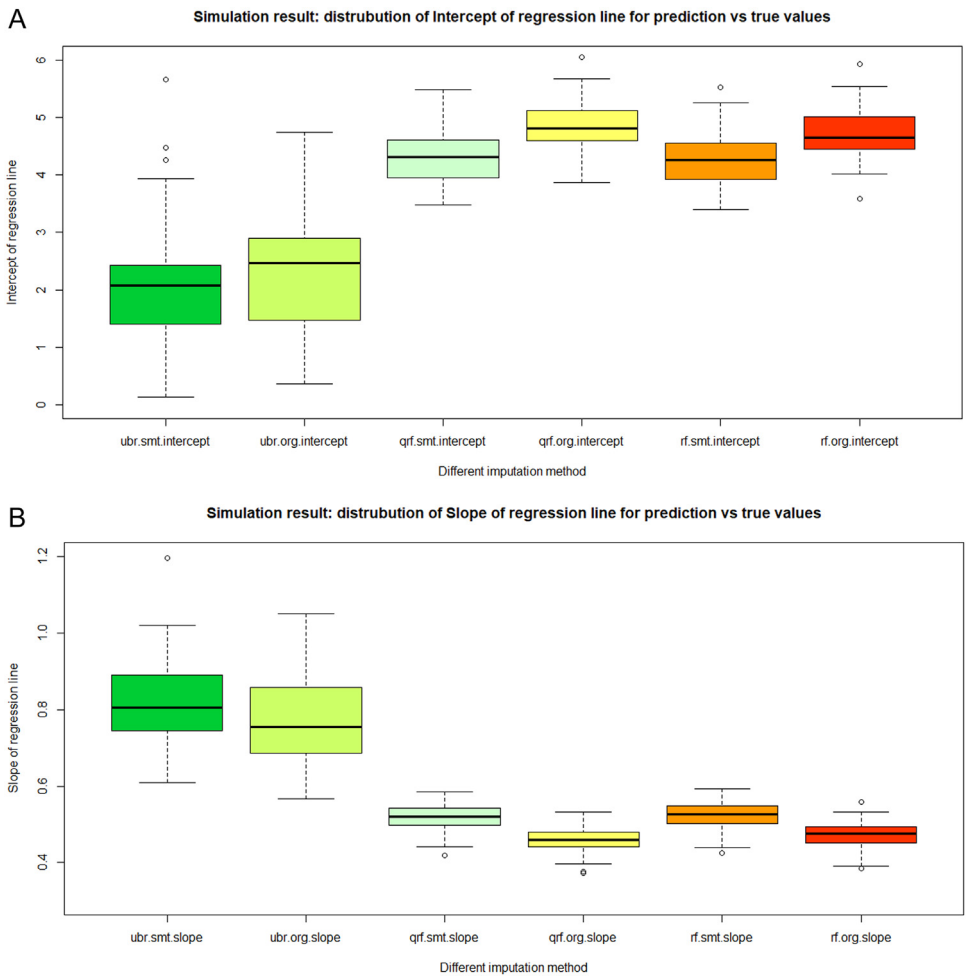

Fig. 8. Simulation result – distribution of the intercepts (A) and slops (B) from simple regression of true data vs. imputed data by different methods. ubr = utility-based regression (coefficient of relevance function = 0.75), smt = SMOTER data, org = original data, qrf = quantile random forests, rf = random forests.

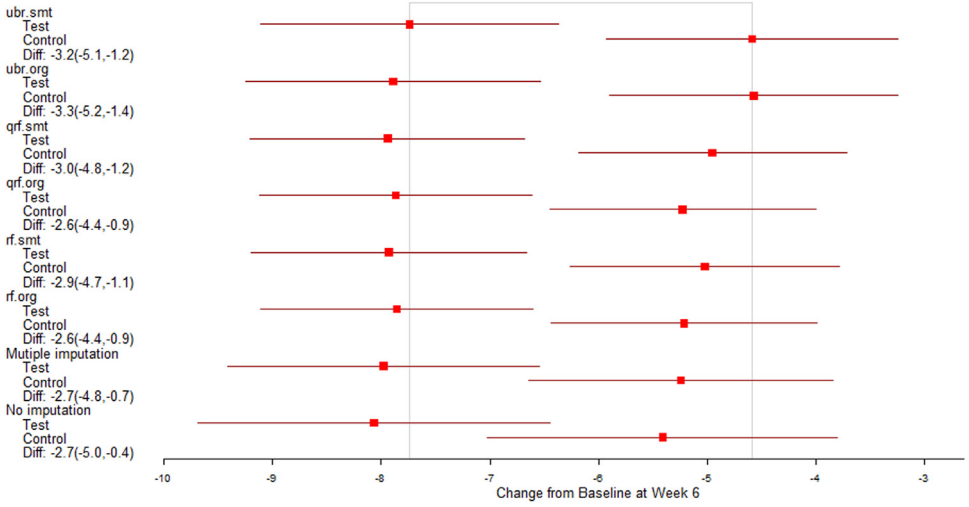

Fig. 9. Real data: forest plot for the analysis results of change from baseline in HAMD17 score at Week 6 using different methods. ubr = utility-based regression, smt = SMOTER data, org = original data, qrf = quantile random forests, rf = random forests.

pecially in the cases when the dropout rate or the efficacy pattern of dropouts are not comparable between treatment group).

5. Discussion

We aim to handle the realistic missing data scenarios (i.e., mixture of MCAR, MAR, and MNAR data) in clinical trials with continuous outcome variable. We treat MNAR as imbalanced learning task. The standard error measures are not suitable for non-unique cost learning. We propose a hybrid imbalanced learning approach that combines UBR with SMOTER. The UBR takes both the prediction error and relevance of the target variable value into account such that the areas been assigned high relevance get more focus

in the learning process. SMOTER is an effective approach to give more weights on the rare cases and also to avoid the model overfitting problem. The relevance function is a crucial part of the proposed method. The choice of the relevance function and its associated parameters should be based on the specific problem in hand and it should be pre-specified. It is inevitable to define the relevance function according to the distribution of available data, and it is also important to make a plausible assumption on the missing data (i.e., the possible locations of target variable scale where the missing data tend to occur) based on the information collected in the clinical trial. We recommend to perform sensitivity analysis with different relevance functions (and associated parameters) to check the appropriateness and robustness of the primary analysis. We evaluate the performance of proposed method in a comprehensive manner in the simulation study. When assessing the impact of missing data on the aggregated estimation, we recommend to evaluate the performance of imputation method not only in terms of the bias (like mean of imputed data) but also in terms of variance the imputed data, which is also an important element in the decision making (e.g., the decision based on the inferential statistics).

The commonly used imputation methods (like random forests and multiple imputation) do not perform as well as the proposed method and showed systematic bias in the aggregated estimation. Those methods tend to underestimate the mean and SD given the presence of MNAR in the area of high values of the target variable. A similar pattern of the systematic bias is also observed in the real data from an antidepressant clinical trial with a dropout rate of 25%. Overall, our hybrid imbalanced learning approach provides plausible prediction for all the MCAR, MAR and MNAR data and reduced the bias of missing data in the aggregated estimation. Therefore, we encourage the integration of utility-based learning strategies for handling of missing data in the analysis of clinical trials.

Limitations of this study include: (1) The use of some specific technical elements, such as QRF and SMOTER, is based on our current knowledge in this domain, and this can be further improved once new and better methods emerge; (2) To demonstrate the basic idea of utility-based regression, we look at the cross-sectional data only. However, in practice, missing data problem is more often in the longitudinal studies. Therefore, from practical point of view, an extension of the utility-based regression in the longitudinal setting is necessary.

Supporting information

All R programs for the whole workflow, datasets and outputs will be available at the website of Computer Methods and Programs in Biomedicine.

Declaration of Competing Interest

The authors have declared no conflict of interest.

Acknowledgements

The authors thank Prof. Anne-Laure Boulesteix for her valuable contribution to this work. The authors also thank the anonymous reviewers, the Associate Editor and the Editor for their generous and constructive detailed comments that helped us to improve the paper.

Appendix

Not applicable.

References

[1] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.
[2] C.K. Enders, in: Applied Missing Data Analysis, Guilford Press, New York, 2010, pp. 295–301. Page.
[3] N.H. Haliduola, F. Bretz, U. Mansmann, Missing data imputation in clinical trials using recurrent neural network facilitated by clustering and oversampling, Biometrical J. 64 (5) (2022) 863–882, doi:10.1002/bimj.202000393.
[4] R.P. Ribeiro, Utility-based Regression, Dep. Computer Science, Faculty of Sciences - University of Porto, 2011 PhD thesis.
[5] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Acad. Ind. Res. 16 (2002) (2002) 321–357.
[6] Torgo, L., Ribeiro, R.P. (2007). Utility-Based Regression. 597–604. 10.1007/978-3-540-74976-9_63.
[7] L. Torgo, R.P. Ribeiro, B. Pfahringer, P. Branco, Smote for regression, in: Progress in Artificial Intelligence, Springer, 2013, pp. 378–389. pages.
[8] F.N. Fritsch, R.E. Carlson, Monotone piecewise cubic interpolation, SIAM J. Numer. Anal. 17 (2) (1980) 238–246.
[9] N. Meinshausen, Quantile Regression Forests, J. Mach. Learn. Res. 7 (2006) 983–999.
[10] Meinshausen, N. (2017). Quantile regression forests, a R package available at https://cran.r-project.org/package=quantregforest.
[11] A.L. Boulesteix, R. Wilson, A. Hapfelmeier, Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies, BMC Med. Res. Methodol. 17 (1) (2017) 138 2017Published 2017 Sep 9, doi:10.1186/s12874-017-0417-2.
[12] A.L. Boulesteix, H. Binder, M. Abrahamowicz, W. Sauerbrei, On the necessity and design of studies comparing statistical methods, Biom J. 60 (2017), doi:10.1002/bimj.201700129.
[13] P. Branco, L. Torgo, R. Ribeiro, A survey of predictive modelling under imbalanced distributions, ACM Comput. Surv. 1 (1) (2016) Article 1, Publication date: January 1.
[14] M.M. Rau, S. Seitz, F. Brimioulle, E. Frank, O. Friedrich, D. Gruen, B. Hoyle, Accurate photometric redshift probability density estimation – method comparison and application, Mon. Not. R. Astron. Soc. 452 (4) (2015) 3710–3725 01 October 2015, Pages, doi:10.1093/mnras/stv1567.
[15] Branco, P., Ribeiro, R.P., Torgo, L. (2017). UBL: an R package for utility-based learning.
[16] L. Breiman, Random Forests, Mach Learn 45 (1) (2001) 5–32, doi:10.1023/A:1010933404324.
[17] A. Liaw, M. Wiener, in: Package "randomForest": Breiman and Cutler's random Forests for Classification and Regression, 4, R Development Core Team, 2018, pp. 6–10.
[18] R. Koenker, Quantile Regression (Econometric Society Monographs), Cambridge University Press, Cambridge, 2005, doi:10.1017/CBO9780511754098.
[19] M.S. Santos, J.P. Soares, P.H. Abreu, H.J. Araujo, Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches, IEEE Comput. Intell. Mag. (2018).
[20] H. Demirtas, B. Doganay, Simultaneous generation of binary and normal data with specified marginal and association structures, J. Biopharm. Stat. 22 (2) (2012) 223–236, doi:10.1080/10543406.2010.521874.
[21] Amatya, A., Demirtas, H., Gao, R. (2020). BinNor: an R package for con-current generation of binary and normal data.
[22] D.B. Rubin, Multiple Imputation For Nonresponse in Surveys, Wiley, New York, 1987.
[23] J. Siddique, T.R. Belin, Multiple imputation using an iterative hot-deck with distance-based donor selection, Stat. Med. 27 (1) (2008) 83–102.
[24] S. van Buuren, K. Groothuis-Oudshoorn, et al., mice: multivariate imputation by chained equations in R, J. Stat. Softw. 45 (3) (2011) 1–67 https://www.jstatsoft.org/v45/i03/.
[25] London School of Hygiene and tropical medicine (2017). (https://missingdata.lshtm.ac.uk/2017/04/28/example-dataset-from-an-antidepressant-clinical-trial/).
[26] D.J. Goldstein, Y. Lu, M.J. Detke, C. Wiltse, C. Mallinckrodt, M.A. Demitrack, Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine, J. Clin. Psychopharmacol. 24 (2004) 389–399.