ORIGINAL ARTICLE

Journal Section

Handling Missing Data in Rheumatoid Arthritis Registry using Random Forest Approach

Ahmad Alsaber[1]* | Adeeba Al-Herz[2]† | Jiazhu Pan[1]* | Ahmad T. AL-Sultan[3]‡ | Divya Mishra[4§] | KRRD Group[2]

[1]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, G1 1XH, UK

[2]Department of Rheumatology, Al-Amiri Hospital, Kuwait

[3]Department of Community Medicine and Behavioural Sciences, Kuwait University, Kuwait

[4]Department of Plant Pathology, Kansas State University, USA

Correspondence
Ahmad Alsaber, Department of Mathematics and Statistics, University of Strathclyde, Glasgow, G1 1XH, UK.
Email: a.alsaber@strath.ac.uk

Funding information
This is an investigator study from KRRD registry. The KRRD registry is supported by unrestricted grants from Pfizer.

Missing data in clinical epidemiological research violate the intention-to-treat principle, reduce the power of statistical analysis, and can introduce bias if the cause of missing data is related to a patient's response to treatment. Multiple imputation (MI) provides a solution to predict the values of missing data. The main objective of this study is to estimate and impute missing values in patient records. The data from Kuwait Registry for Rheumatic Diseases (KRRD) was used to deal with missing values among the patients records. A number of methods were implemented to deal with missing data, however choosing the best imputation method was judged by the lowest root mean square error (RMSE). Among 1735 rheumatoid arthritis (RA) patients, we found missing values vary from 5% to 65.5% of the total observations. The results show that sequential random forest method can estimate these missing values with a high level of accuracy. The RMSE varied between 2.5 and 5.0. MissForest had the lowest imputation error for both continuous and categorical variables under each missing data rate (10%, 20%, and 30%) and had the smallest prediction error difference when the models used the imputed laboratory values.

Abbreviations: KRRD, Kuwait Registry for Rheumatic Diseases; kNN, k-nearest neighbors; RMSE, Root Mean Square Error.

* Equally contributing authors.

KEYWORDS
rheumatoid, missing values, imputation techniques , random forest, *kNN*, *KRRD*

1 | INTRODUCTION

In any clinical research, missing values or experimental values remain a problem in correctly analyzing results and obtaining inaccurate outcomes. These missing values often lead to misinterpretation and biased results, which could ultimately affect the overall conclusion of an investigation [1, 2, 3, 4]. The application of statistical analyses in experiments with missing values poses serious problems, as the missing values are often automatically ignored by the statistical algorithms. The results obtained by the investigator in such experiments may be insignificant or even meaningless [5, 6, 7]. Missing data is a common problem for all kind of research data, especially in clinical trials. It always becomes problematic when sample collection was not performed in random order or were obtained using an improper methodology [8]. Certain factors are responsible for missing values in the data of a study: (i) the data are not captured due to some unknown reason, such as error in recording the data from an electronic detector/data recorder or manual recording by technical medical staff; (ii) data are missing due to a known reason, such as critical medical conditions; or (iii) data are not recorded as they are unrelated to the patient's clinical medical condition [6]. However, the biased and misleading information obtained when values are missing can be managed by the application of imputation methods.

1.1 | Missing Imputation - Rubin's Approach

Imputation involves the substitution of missing values with known variables. This type of approach is widely used, as it produces complete data. However, the decision regarding the imputed value cannot be unbiased (e.g. Multiple imputation for missing data makes it possible for the researcher to obtain approximately unbiased estimates of all the parameters from the random error. The researcher cannot achieve this result from deterministic imputation, which the multiple imputation for missing data can do), as it could lead to an overestimation of confidence in the outcome. To overcome this problem, Rubin suggested the theory of multiple imputation, in which missing values are imputed using the appropriate model a few times (generally three to five times) and a standard method is applied for the analysis [9, 10, 11, 4]. The imputation method provides more accurate results, but problems with the application of imputation include: (1) maximum use of the available data to reduce the error for univariate data and preserve covariance in multivariate data sets; and (2) reporting the variance estimates of uncertainty caused due to the imputed value [11]. Several parametric and non-parametric techniques have been employed to deal with missing values. Parametric methods depend on the assumed method, whereas non-parametric methods require a high number of observations [12].

1.2 | Categorization of Missing Values

Rubin categorized the missing value problem into three groups: missing completely at random (MCAR), missing at random (MAR), and not missing at random/missing not at random (NMAR/MNAR) [10, 12, 13]. The MAR method is generally used in clinical epidemiological research [14, 15]. It is critical to determine the category of the data, in order to choose a statistical strategy [16, 17].

1.2.1 | Missing at Random (MAR):

Data are MAR if the missingness depends on the observed characteristics, not the unobserved characteristics, meaning that the relationships observed in the data can be used to predict the occurrence of missing values.

1.2.2 | Missing Completely at Random (MCAR):

As the randomness of MAR is conditional on observed characteristics, which distinguishes it from the completely-at-random type of MCAR, dropping or omitting those cases with missing values in the analysis may lead to biased results [15].

1.2.3 | Missing not at Random (MNAR):

Data are considered MNAR if their missingness depends on characteristics that are not observed and cannot be fully explained by the observed characteristics. Systematic differences between missing and non-missing data exist for data that are MNAR. In some circumstances, randomness in the missing data mechanism may be ignored without affecting the inference [10]. Both MCAR and MAR can be considered ignorable, in the sense that a proper method (e.g., multiple imputation) may recover the missing information without modeling. In contrast, the MNAR mechanism requires a method that considers the missing data mechanism to make inferences about the complete (and partially unobserved) data; in other words, the model for the missing data mechanism cannot be ignored [18].

1.3 | Methods used for imputing missing values

To treat and estimate the missing values, [19] proposed a non-parametric random forest (RF) model, which is an extended version of classification and regression trees (CARTs) and involves a supervised learning group method. The method used to build the trees involves replacement sampling of the main data set. The classification and regression trees are created using the training data bootstrap samples and tree induction using random feature selection [20, 21]. The performance of a tree is evaluated on the remaining data, which are contained in an out-of-bag sample.

1.3.1 | Missing imputation using Random Forest (RF):

The best RF is determined based on the out-of-bag error, which is an unprejudiced gauge of the true prediction error [22]. RF has following advantages: (1) it is applicable even when number of variables is greater than the number of samples; (2) it is not prone to multicollinearity; (3) it is suitable for non-linear trends; (4) it does not suffer from the overfitting problem with an increase in the number of trees; and (5) it can tolerate outliers and missing values [23].

1.3.2 | Missing imputation using sequential random forest (missForest):

Another algorithm based on RF, called sequential random forest (missForest), has recently been developed for missing data imputation [24]. This algorithm can impute missing values on any kind of data and its goal is the prediction of every single missing value, instead of drawing random values from a distribution. This algorithm can handle multivariate data sets concurrently comprising categorical and continuous variables [25]. The key advantages of missForest over other imputation methods include: (1) having no requirement for the tuning of parameters; (2) it does not depend

on assumptions pertaining to the distribution of data sets; (3) it allows for assessment of imputation quality without setting test data or laborious cross-validations using out-of-bag imputation error estimates; and (4) it provides above-par imputation results, even for high-dimensional data sets (i.e., when the number of variables is greater than the number of observations) [24].

The missForest approach is based on a decision tree that is supervised by a machine learning algorithm that can be used for both classification and regression problems. A decision tree is simply a series of sequential decisions made to reach a specific result. It initially imputes all missing data using the mean/mode, then for each variable with missing values, MissForest fits a random forest on the observed part and then predicts the missing part (sequential movement). This process of training and predicting repeats in an iterative process until a stopping criterion is met, or a maximum number of user-specified iterations is reached. The reason for the multiple iterations is that, from iteration two onwards, the random forests performing the imputation will be trained on better quality data that itself has been predictively imputed. In other words, the method treats the variable of the missing value as a predictor and borrows information from other variables by the resampling-based classification and regression trees to grow a random forest for the final prediction. The method is repeated until the imputed values reach convergence [26, 27].

In general, missForest and k-NN are considered as machine learning algorithms because they do not explicitly require the users to define how the prediction is taking place, whereas multiple imputation and seasonal decomposition require model specifications by the users.

In the present investigation, we consider four data mining techniques to predict the missing values in the Kuwait Registry for Rheumatic Diseases (KRRD): predictive mean matching (PMM), k-nearest neighbors (kNN), random forest (RF), and sequential RF (missForest). The main objective of this study was to handle missing data in the KRRD, where the amount of missing data varied between 1% and 65.5% (Table 1). Our secondary objectives were to choose the best missing data mechanism (MAR, MCAR, or MNAR) when assuming three different rates of missingness (10%, 20%, and 30%), as well as to compare the selected imputation methods (PMM, RF, kNN, and missForest) for each missing data mechanism under each rate of missingness. To select the best method for imputing missing data in KRRD, the root mean square error (RMSE) was used to evaluate the best imputation method which minimized the difference between the imputed data points and the original data points (that were subsequently set to missing).

2 | METHODS

2.1 | Data Source—Kuwait Registry for Rheumatic Diseases (KRRD)

All rheumatoid arthritis (RA) patients in this study were officially registered in the Kuwait Registry for Rheumatic Diseases (KRRD). The KRRD is a national registry listing adult patients with rheumatic diseases. Patients who fulfilled the American College of Rheumatology (ACR) criteria for RA [28] registered from January 2012 through March 2020 were included in the study. The RA information data were collected from the rheumatology departments of four major government hospitals in Kuwait, based on patient visits. The selected hospitals are mainly distributed in different governorates covering the ethnic diversity of the Kuwaiti population. The KRRD, from which this study originated, was approved by the Ethics Committees of the Faculty of Medicine at Kuwait University and the Ministry of Health. Additionally, informed consent was obtained from all represented patients enrolled in the registry [29].

Using the data obtained from KRRD, we conducted an in-depth comparative analysis of the different imputation methods. Missing data were entered into each data set, assuming a general missing data pattern and three mechanisms of missing data: MCAR, MAR, and NMAR. Under the MCAR assumption, missing values were randomly applied to each data set. Under the MAR assumption, the probability of information being missing depended on class attribute.

TABLE 1 Study variables with abbreviations.

This table provides the frequencies

Variable name	Abbreviation	Measures	Type of variable	Missing rate	Variable role
RA Disease Duration		baseline	scale	12.4%	independent
Smoking		baseline	categorical	26.0%	independent
Rheumatoid Factor	RF	baseline	categorical	8.3%	independent
Antinuclear Antibodies	ANA	baseline	categorical	21.4%	independent
Anti-Cyclic Citrullinated Peptide	ACPA	baseline	categorical	21.0%	independent
Sicca Symptoms	SICCA	baseline	categorical	19.8%	independent
Rheumatoid Nodules	Nodules	baseline	categorical	18.5%	independent
Family hHstory of Rheumatic Disease	FH	baseline	categorical	28.4%	independent
Treatment Class	TC	repeated	categorical	13.7%	independent
Steroid Therapy	Steroid	baseline	categorical	6.6%	independent
Joint Pain		repeated	categorical	3.8%	independent
Disease Activity Score 28	DAS28	repeated	scale	1.0%	target (outcome)
Erythrocyte Sedimentation Rate	ESR	repeated	scale	5.1%	independent
C-Reactive Protein	CRP	repeated	scale	2.2%	independent
Health Assessment Questionnaire Disability Index	HAQ	repeated	scale	65.5%	independent

Under the NMAR assumption, the largest or smallest values of X_s were removed. The objective of the study was to derive a comparison of four different imputation methods for NMAR, MAR, and MCAR, concerning missing data. We simulated the rates of missing data by varying the value proportions by 10%, 20% and 30%.

2.2 | Calculating RA Indices

RA disease activity scores are measured using two different indices: DAS28 and CDAI. The DAS28 is the sum of four outcome parameters: TJC28, the number of tender joints (0–28); SJC28, the number of swollen joints (0–28); ESR, the erythrocyte sedimentation rate (in mm/h) (C-reactive protein (CRP) may be used as an alternative to ESR in the calculation); and GH, the patient global health assessment (from 0 = best to 100 = worst) (Equation 1).

$$\text{DAS- } 28 = 0.56 \times \sqrt{TJC28} + 0.28 \times \sqrt{SJC28} + 0.70 \times \ln(ESR \quad Or \quad CRP) + 0.014 \times GH. \quad (1)$$

The second index is the Clinical Disease Activity Index (CDAI). THe CDAI considers the following items: TJC28, the number of tender joints (0–28); SJC28, the number of swollen joints (0–28); PaGH, the patient global health assessment (from 0 = best to 10 = worst); and PrGH, the care provider global health assessment (from 0 = best to 10 = worst) (see Equation 1). In this study, we used the first index (DAS28) as a target variable.

2.3 | Multiple Imputation (MI) Process Using Rubin's Rules

For our data sets, we used Rubin's rules [10] for handling missing data. The MI process was conducted separately for each variable in the data set (Figure 1). The first step in multiple imputation is to create values ($imputes$ or m_i),

with 5 iterations for each m_i (Imputed: set 1 to set 5, see figure 1) in to be substituted for the missing data. To create the imputed values, we need to identify a model (e.g., a linear regression) that allows us to create imputes based on other variables in the data set (predictor variables). As we needed to perform this multiple times to produce multiple-imputed data sets, we identified a set of regression lines that were similar to each other.

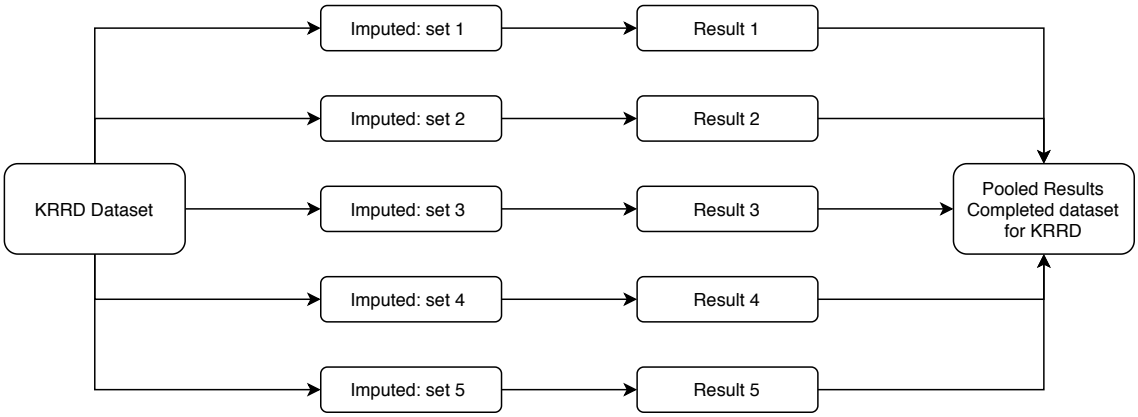

FIGURE 1 The steps of implementing multiple imputations using Rubin's rules to estimate missing values for KRRD.

2.4 | Number of Needed Imputations

An important aspect of previous technical treatments of multiple imputation is that the discussion of selecting the number of imputations that are required for acceptable statistical inference (e.g., [30, 31, 32]). For example, Schafer and Olsen [32] recommend that in several applications, simply 3–5 imputations are enough to get sufficient results. Many are surprised by the claim that only 3–5 imputations may be needed. Rubin [30] shows that the efficiency of an estimate based on m imputations is approximately

$$\left(1+\frac{\gamma}{m}\right)^{-1} \tag{2}$$

where γ is the fraction of missing information for the quantity being estimated gains rapidly diminish after the first few imputations. In most situations there's merely very little advantage to generate and analyzing over a few imputed datasets. In theory, the more imputation, the better performance in estimating missing values, but it takes a lot of time, which is a barrier for this research. It is convenient to set m = 5 during the stage of model building and raising the amount in the evaluation stage if it is needed [33]. So, in this study, The MI methods are performed with $M = 5$ imputed data sets which can be considered as satisfactory [30].

2.5 | Multiple Imputation Using RF Method

Assume that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$ is an $n \times p$-dimensional data matrix. We propose using the random forest technique to impute missing observations. The random forest algorithm has a built-in routine to handle the values that are missing

by weighing the frequencies of values with the proximity of a random forest after the training of the mean data set is initially imputed [34]. This approach needs a response variable that is complete and useful for forest training. Instead, we directly estimate the values of all the missing values using a random forest that is trained from the observed data set, where X is the matrix of the complete data. $\mathbf{X}_s$ contains all missing values at entries $\mathbf{i}_{mis}^{(s)} \subseteq \{1, \ldots, n\}$. The data set can be separated into four parts:

1. $\mathbf{y}_{obs}^{(s)}$: the observed values of $\mathbf{X}_s$;
2. $\mathbf{y}_{mis}^{(s)}$: the missing values of $\mathbf{X}_s$;
3. $\mathbf{x}_{obs}^{(s)}$: the observations $\mathbf{i}_{obs}^{(s)} = \{1, \ldots, n\} \backslash \mathbf{i}_{mis}^{(s)}$, which belong in the other variables $\mathbf{X}_s$;
4. $\mathbf{x}_{mis}^{(s)}$: the observations $\mathbf{i}_{mis}^{(s)}$ that belong in the other variables $\mathbf{X}_s$.

Note that $\mathbf{X}_{obs}^{(s)}$ and $\mathbf{X}_{\text{mis}}^{(s)}$ are not completely observed, as the index $\mathbf{i}_{obs}^{(s)}$ corresponds to the observed values of the variable $\mathbf{X}_s$.

According to [24], the process starts with an initial guess for missing values in $\mathbf{X}$ using a mean imputation approach or any other imputation method depending on the data. Then, we sort the predictors $\mathbf{X}_s$, $s = 1, \ldots, p$, in ascending or descending order, $\mathbf{X}_s$, $s = 1, \ldots, p$, according to the number of missing values. Then, for each variable $\mathbf{X}_s$, the missing values are imputed by random forest (i.e., the first fitting) with response $\mathbf{y}_{obs}^{(s)}$ and predictors $\mathbf{X}_{obs}^{(s)}$. Next, the missing values $\mathbf{y}_{\text{mis}}^{(s)}$ are estimated by adopting the trained random forest to $\mathbf{X}_{\text{mis}}^{(s)}$. The imputation approach should be repeated until a stopping criterion is reached. Pseudo Algorithm 1 shows a representation of the missForest method (Algorithm 1). The stopping criterion (γ) is met when the difference between the last imputed data matrix and the previous one increases for the first time with respect to both variable types. Here, the difference for the set of continuous variables $\mathbf{N}$ is defined as:

$$\Delta_N = \frac{\sum_{j \in \mathbf{N}} \left(\mathbf{X}_{new}^{imp} - \mathbf{X}_{old}^{imp} \right)^2}{\sum_{j \in \mathbf{N}} \left(\mathbf{X}_{new}^{imp} \right)^2}, \tag{3}$$

and that for the set of categorical variables $\mathbf{F}$ is defined as:

$$\Delta_F = \frac{\sum_{j \in \mathbf{F}} \sum_{i=1}^{n} \mathbf{I}_{\mathbf{X}_{new}^{imp} \neq \mathbf{X}_{old}^{imp}}}{\#NA}, \tag{4}$$

where $\mathbf{X}$ is an $n \times p$ matrix, setting the stopping criterion (γ) and initial guesses for missing values; $\mathbf{k} \leftarrow$ is the vector of sorted indices of columns in $\mathbf{X}$ with respect to increasing the amount of missing values; and $\mathbf{X}_{old}^{imp}$ stores the previously imputed matrix. We fit a random forest $\mathbf{y}_{obs}^{(s)} \sim \mathbf{x}_{obs}^{(s)}$; predict $\mathbf{y}_{mis}^{(s)}$ using $\mathbf{x}_{mis}^{(s)}$; use $\mathbf{X}_{new}^{imp}$ to update the imputed matrix using the predicted $\mathbf{y}_{mis}^{(s)}$; and update γ using the imputed matrix $\mathbf{X}^{imp}$, where NA is the number of missing values in the categorical variables.

After imputing the missing values, the performance of different methods was assessed using the normalized root mean squared error (NRMSE) [35] for the continuous variables, defined by:

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\left(\left(\mathbf{X}^{\text{true}} - \mathbf{X}^{\text{imp}} \right)^2 \right)}{\text{var}\left(\mathbf{X}^{\text{true}} \right)}}, \tag{5}$$

Algorithm 1 Impute missing values with random forest [24].

Require: $\mathbf{X}$ is an $n \times p$ matrix. Set up the stopping criterion (γ)

set up initial guess for missing values;
$\mathbf{k}$ is the vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing the amount of missing values;
while not γ **do**
  $\mathbf{X}_{old}^{imp}$ stores the previously imputed matrix;
  for s in $\mathbf{k}$ **do**
    Fit a random forest: $\mathbf{y}_{obs}^{(s)} \sim \mathbf{x}_{obs}^{(s)}$;
    Predict $\mathbf{y}_{mis}^{(s)}$ using $\mathbf{x}_{mis}^{(s)}$;
    $\mathbf{X}_{new}^{imp}$ updates the imputed matrix using predicted $\mathbf{y}_{mis}^{(s)}$;
  end for
  update γ
end while
return the imputed matrix $\mathbf{X}^{imp}$

where $\mathbf{X}^{\text{true}}$ and $\boldsymbol{X}^{imp}$ are the complete data matrix and the imputed data matrix, respectively. In this study, all predictors were classified as continuous observations. The mean and variance are used as a short notation for the empirical mean and variance computed over the missing values only, respectively. When an RF fits to the part that is observed on a variable, we reach the out-of-bag (OOB) estimate of the error for the variable. When the stopping criterion (γ) is met, we average it over the variable set of that type to obtain an approximation of the actual errors of imputation. We assessed the estimation performance by comparing the absolute difference between the OOB imputation error estimate in all simulation runs and the true imputation error.

2.6 | Evaluation Criteria

To determine the best imputation method, three model performance tests were considered [36]: root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (R), which are respectively calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{7}$$

where y_i and $\hat{y}_i$ are the i^{th} observations for the reconstructed and comparison data sets, respectively. The error is measured based on the difference between the estimated and observed values. For RMSE and MAE, the smaller the value obtained, the more accurate the estimation method.

3 | RESULTS

The performance of three imputation mechanisms (MCAR, MAR, and MNAR) was analyzed using sub-data sets from KRRD patients using three different missingness rates (10%, 20%, and 30%). A total of 1735 patients (62.8% men and 37.2% women) from 2012–2020 were included in this study (Table 2). The baseline investigated patient characteristics included factors such as smoking, RF, SICCA, ANA, ACPA, family history, treatment class, comorbidity, steroid and joint pain.

The average duration of RA disease was 9.19 ± 6.76 (SD) years. Most of the data were recorded at Amiri and Farwaniya hospitals (79%). A majority of the patients were non-smokers (89.6%), 77.1% were RF-positive, 65.9% of RA patients were ACPA-positive, and 58.8% had joint pain. The results showed that RA patients had positive SICCA (18.7%), positive ANA (28.4%), positive family history (18.4%), and positive steroid use (22.6%). Table 3 provides a descriptive analysis of RA lab tests for ESR, CRP, HAQ, and DAS28, which were calculated five different times from five different data sets after implementing missing values methods (original data set compared with imputed data sets using PMM, RF, kNN, and missForest).

The mean and SD value of ESR, CRP, HAQ, and DAS28 were 27.5729 ± 22.2706, 5.9904 ± 4.9334, 0.9517 ± 0.6649, and 2.6756 ± 1.2902, respectively, for the original data, and those for the imputed data sets values ranged between 27.0639 and 27.1245, 6.4323 and 6.4456, 0.9042 and 0.9062, and 2.6759 and 2.6767, respectively. The skewness and kurtosis values were mostly positive: 1.2248, 1.0182, 0.8120, and 0.5963 for skewness and 1.6256, 0.3932, −0.0311, and 0.2599 for kurtosis in the case of imputed data sets that ranged between 1.2551 and 1.2675, 0.8085 and 0.8135, 1.1352 and 1.1430, and 0.5961 and 0.6005 for skewness, respectively, and 1.7394 and 1.7854, 0.1395 and 0.1493, 2.1178 and 2.1560, and 0.2664 and 0.2798 for kurtosis, respectively. The data showed that the original data set and the imputed data sets had very close values with small differences for all RA lab tests (ESR, CRP, HAQ, and DAS28).

3.1 | Predicting the Influence of RA Factors on DAS28 Using the Original Data Set

Here we are trying to estimate a regression model that explaining the effect from the independents variables (see Table 1) toward the outcome variable (DAS28). We used the original Kuwait Registry for Rheumatic Diseases (KRRD) dataset. As we mentioned before, the original dataset contains missing values in all variables (see Table 1).

The missing values rate in the original data set are vary from 2% to 66%. Table 4 shows the estimated parameters that predicting DAS28 using a multiple linear model. Only six variables were found to be a significant risk factors that influancing DAS28 ($R^2_{DAS28} = 0.773$). The results showed that ESR, CRP, HAQ, disease duration, and current steroid use were risk factors predicting DAS28, with β = 0.034, 0.020, 0.129, 1.489, 0.247, 1.095, respectively (95% CIs: 0.032–0.036, 0.012–0.029, 0.075–0.183, 1.415–1.562, 0.138–0.356, and 0.963–1.228, respectively). Other factors (RF, ANA, ACPA, SICCA, nodules, smoking, family history and joint pain) were not found to be risk factors influencing DAS28 (Table 4).

Because of the existing of missing values, smoking, joint pain and SICCA were not found to be a significant risk factor for DAS28. However, many scholars showed that those variables (smoking, joint pain and SICCA) can be a risk factors toward DAS28 (e.g. [37], [38] and [39])

3.2 | Predicting the Influence of RA Factors on DAS28 from the PMM-Imputed Data Sets

Using the imputation process to predict all missing values in the KRRD data set using the three different missing imputation mechanisms (MAR, MCAR, and MNAR), we constructed a quality data set after fixing all missing values using PMM. Table 5 shows the estimated parameters when predicting DAS28 using multiple linear models and the PMM-imputed data sets.

The regression results showed the same significant risk factors, this time adding RF to predict DAS28 ($R^2_{DAS28} = 0.727$); in addition, smoking was found to be a very strong risk factor when we used the imputed data sets, but was not when we used the original data set to predict the influence on DAS28.

The regression model that used the original data set did not indicate that RF has a significant influence on DAS28 but, if we used the PMM-imputed data set, the regression model indicated that RF had a significant influence when predicting DAS28.

Due to the effect of bias induced by the missing values, the RF results were not significant when using the original data set; however, after we imputed all the missing data in the KRRD data set, the regression results became more sufficient and reliable.

3.3 | Predicting the Influence of RA Factors on DAS28 from the Imputed Data Sets Using kNN

The kNN-based imputation of missing data restored the leftover values in the KRRD data set and better standard data were obtained. As shown in Table 6, various parameters were used to establish DAS28 prediction using the kNN-imputed data sets.

The regression results were calculated based on similar factors as those of the original data set with further incorporation of RF and SICCA and further prediction of DAS28 ($R^2_{DAS28} = 0.727$).

Similar to the PMM-based analysis, the kNN imputation revealed that RF can significantly influence the prediction of DAS28, which was not significant in the original data set due to bias. The disease duration factor had a negative value in both PMM and kNN, whereas it was positive in the original data set.

3.4 | Predicting the Influence of RA Factors on DAS28 from the RF-Imputed Data Sets

The results of RF imputation, in terms of removing bias in the KRRD data set, were similar to those of PMM-based imputation. Table 5 shows that the factors that significantly affected the DAS28 parameter at $p < 0.01$ were similar between PMM- and RF-based imputation, with their values being very close. The adjusted R^2_{DAS28} value of 0.728 was obtained after imputation.

3.5 | Predicting the Influence of RA Factors on DAS28 from the missForest-Imputed Data Sets

One of the best methods for imputation reported in the literature and evident from the analysis was missForest. Apart from all the factors listed in the original data set and compared to the imputation by PMM, kNN, and RF, the missForest-based imputation analysis produced better results, as evidenced by MAR, MCAR, and MNAR missing value mechanisms (Table 7 and Table 5).

The adjusted R^2_{DAS28} value was 0.731. The data set was the most refined and its quality was the most improved

after applying the missForest imputation method.

We hypothesized that the missing data could be imputed using the different imputation strategies; therefore, the MCAR, MAR, and MNAR mechanisms were simulated for missing values in the three different missingness proportions of 10%, 20%, and 30%. To avoid bias in the comparison, we used four multiple imputation methods: PMM, kNN, RF, and missForest (Table 7).

The KRRD data set was simulated with these imputation methods; the best method was selected according to the RMSE score. As shown in Table 7, the RMSE value ranged between 2.518 to 6.066 for MAR, 2.555 to 5.590 for MCAR, and 3.631 to 8.004 for MNAR. The MAR had the lowest RMSE, compared to the other missing data methods.

Similar investigations have been previously performed and our data were in agreement with those in the earlier reports [40, 41].

In the MAR, MCAR, and MNAR mechanisms, missForest was the best method of imputation, having the lowest RMSE values for all of the parameters and at all three percentages of simulated missing data (MAR: 2.518, 3.013, and 3.032; MCAR: 3.168, 2.555, and 2.871; and MNAR: 4.962, 4.180, and 3.631 for 10%, 20%, and 30%, respectively) this results agreed with [42]; [43]; [40] and [24].

This was followed by kNN, which performed better than the other two imputation methods (RF and PMM), in terms of RMSE values, at every percentage of missingness (MAR: 4.107, 4.884, and 4.184; MCAR: 3.820, 3.560, and 3.734; and MNAR: 6.236, 5.507, and 5.062 for 10%, 20%, and 30%, respectively); see Table 7. Similar results have been reported that strongly support the better imputation of kNN, compared with RF and PMM ([44]). RF and PMM were the worst-performing multiple imputation methods; of the two, using RF had a slight advantage over PMM but PMM had better imputation in a few of the cases, such as MAR 30% or MNAR 10% and 30%, where RF had a larger RMSE value than PMM. Table 5 and Table 6 represent the multiple regression coefficients with 95% confidence intervals (CIs) for the prediction of DAS28 using the imputed data sets (PMM, RF, kNN, and missForest). The table demonstrates the effect of patient demographics on RA disease activity, where DAS28 was the response variable.

The disease activity score for DAS28 is also reported; where R^2_{DAS28} = 0.727 for PMM method, and R^2_{DAS28} = 0.728 for RF method. Regarding kNN and missForest, R^2_{DAS28} = 0.728 for kNN method, and R^2_{DAS28} = 0.731 for missForest method. The results depict the positive effect of various factors, such as ESR, CRP, HAQ, RF, SICCA, smoking, joint pain, and current steroid use, with β = 0.031, 0.015, 0.202–0.204, 0.050–0.061, 0.057–0.065, 0.131–0.140, 0.674–0.677, and 0.118–0.129, respectively, on RA disease activity, whereas family history and disease duration—with β = (0.029) to (–0.021) and â0.007, respectively—had negative effects under all four imputation methods (Table 5 and Table 6).

Additionally, nodules and constant showed diverse effects per imputation method. The nodules had positive values for PMM and missForest, with β = 0.016 and 0.004, respectively, and negative values for RF and kNN, with β = 0.011 and 0.002, respectively. Constant had negative values for PMM, kNN, and missForest, with β = 0.032, 0.017, and 0.037, respectively, and a positive value for RF, with β = 0.002 (Table 5 and Table 6).

TABLE 2 Baseline patient characteristics of KRRD (2012 to 2020).

	[ALL]	N
	N=1735	
Sex (Female)	1090 (62.8%)	1735
Age (years)	54.0 (12.6)	1719
RA Disease Duration (years)	9.19 (6.76)	1520
Nationality		1735
Kuwaiti	839 (48.4%)	
Non-Kuwaiti	896 (51.6%)	
Main Hospital		1735
Amiri	708 (40.8%)	
Farwaniya	663 (38.2%)	
Jahra	83 (4.78%)	
Mubarak	280 (16.1%)	
Sabah	1 (0.06%)	
Smoking (Yes)	133 (10.4%)	1284
RF (Positive)	1227 (77.1%)	1591
SICCA (Yes)	260 (18.7%)	1391
ANA (Positive)	388 (28.4%)	1364
ACPA (Positive)	903 (65.9%)	1370
Family History (Positive)	229 (18.4%)	1243
Treatment Class (Biologics)	488 (32.6%)	1498
Co-morbidity (Yes)	926 (53.4%)	1735
Current Steroid (Yes)	366 (22.6%)	1620
Joint Pain (Yes)	982 (58.8%)	1670

TABLE 3 The mean and standard deviation for ESR, CRP, HAQ, and DAS28 from the original data set and the imputed data sets (IM).

Data Set	Variable	*N*	Minimum	Maximum	Mean	SE	SD	Skewness	Kurtosis
Original data	ESR	10703	0.0000	134.0000	27.5729	0.2153	22.2706	1.2248	1.6256
	CRP	8769	0.0000	21.0000	5.9904	0.0527	4.9334	1.0182	0.3932
	HAQ	4004	0.0125	3.0000	0.9517	0.0105	0.6649	0.8120	−0.0311
	DAS28	11213	0.0000	9.7050	2.6756	0.0122	1.2902	0.5963	0.2599
$IM_1 = PMM$	ESR	11282	0.0000	134.0000	27.0701	0.2066	21.9426	1.2589	1.7503
	CRP	11282	0.0000	21.0000	6.4456	0.0441	4.6873	0.8108	0.1489
	HAQ	11282	0.0125	3.0000	0.9053	0.0044	0.4647	1.1356	2.1236
	DAS28	11282	0.0000	9.7050	2.6761	0.0121	1.2883	0.6005	0.2798
$IM_2 = RF$	ESR	11282	0.0000	134.0000	27.0639	0.2068	21.9637	1.2675	1.7854
	CRP	11282	0.0000	21.0000	6.4426	0.0442	4.6961	0.8135	0.1493
	HAQ	11282	0.0125	3.0000	0.9042	0.0044	0.4657	1.1352	2.1178
	DAS28	11282	0.0000	9.7050	2.6763	0.0121	1.2878	0.5992	0.2717
$IM_3 = kNN$	ESR	11282	0.0000	134.0000	27.1245	0.2074	22.0287	1.2609	1.7410
	CRP	11282	0.0000	21.0000	6.4323	0.0441	4.6865	0.8085	0.1395
	HAQ	11282	0.0125	3.0000	0.9061	0.0044	0.4658	1.1430	2.1560
	DAS28	11282	0.0000	9.7050	2.6767	0.0121	1.2876	0.5974	0.2689
$IM_4 = missForest$	ESR	11282	0.0000	134.0000	27.0939	0.2064	21.9236	1.2551	1.7394
	CRP	11282	0.0000	21.0000	6.4396	0.0440	4.6772	0.8088	0.1444
	HAQ	11282	0.0125	3.0000	0.9062	0.0044	0.4628	1.1375	2.1519
	DAS28	11282	0.0000	9.7050	2.6759	0.0121	1.2861	0.5961	0.2664

TABLE 4 Multiple regression coefficients with 95% confidence intervals (in parentheses) for predicting DAS28 using the original data set including the missing values.

	DAS28
	Data Set = Original
ESR	0.034*** (0.032, 0.036)
CRP	0.020*** (0.012, 0.029)
HAQ	0.129*** (0.075, 0.183)
RF	0.021 (−0.064, 0.106)
ANA	−0.062 (−0.139, 0.014)
ACPA	0.008 (−0.066, 0.082)
SICCA	0.083 (−0.011, 0.178)
Nodules	−0.519 (−1.122, 0.083)
Smoking	0.184 (-0.019, 0.350)
Family History	−0.087 (−0.174, 0.001)
Joint Pain	0.273 (-0.015, 0.530)
Disease Duration	1.489*** (1.415, 1.562)
Current Steroid	0.247*** (0.138, 0.356)
Constant	1.095*** (0.963, 1.228)
R^2	0.773
Adjusted R^2	0.769
Note:	$^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

TABLE 5 Multiple regression coefficients with 95% confidence intervals (in parentheses) to predict DAS28 from other predictors from PMM- and RF-imputed data sets.

	DAS28	
	Imputed data set	
	(PMM)	(RF)
ESR	0.031*** (0.030, 0.031)	0.031*** (0.030, 0.031)
CRP	0.015*** (0.012, 0.017)	0.015*** (0.012, 0.017)
HAQ	0.202*** (0.178, 0.226)	0.202*** (0.178, 0.225)
RF	0.061*** (0.035, 0.087)	0.050*** (0.024, 0.076)
ANA	0.005 (−0.020, 0.031)	0.003 (−0.023, 0.028)
ACPA	0.003 (−0.020, 0.026)	0.008 (−0.016, 0.031)
SICCA	0.064*** (0.034, 0.094)	0.060*** (0.031, 0.090)
Nodules	0.016 (−0.044, 0.077)	−0.011 (−0.071, 0.049)
Smoking	0.131*** (0.086, 0.177)	0.140*** (0.095, 0.186)
Family History	−0.029 (−0.059, 0.001)	−0.022 (−0.052, 0.008)
Joint Pain	0.674*** (0.662, 0.686)	0.676*** (0.664, 0.688)
Disease Duration	−0.007*** (−0.009, −0.006)	−0.007*** (−0.009, −0.005)
Current Steroid	0.128*** (0.093, 0.162)	0.118*** (0.084, 0.153)
Constant	−0.032 (−0.146, 0.083)	0.002 (−0.112, 0.117)
Observations	11,282	11,282
R^2	0.727	0.728
Adjusted R^2	0.727	0.728
Note:		$^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

TABLE 6 Multiple regression coefficients with 95% confidence intervals (in parentheses) to predict DAS28 from other predictors from kNN- and missForest-imputed data sets.

	DAS28	
	Imputed data set	
	(kNN)	(missForest)
ESR	0.031*** (0.030, 0.031)	0.031*** (0.030, 0.031)
CRP	0.015*** (0.013, 0.018)	0.015*** (0.012, 0.017)
HAQ	0.203*** (0.180, 0.227)	0.204*** (0.180, 0.227)
RF	0.058*** (0.032, 0.084)	0.056*** (0.030, 0.082)
ANA	0.001 (−0.025, 0.026)	0.008 (−0.018, 0.033)
ACPA	0.004 (−0.019, 0.027)	0.004 (−0.020, 0.027)
SICCA	0.065*** (0.035, 0.094)	0.057*** (0.027, 0.087)
Nodules	−0.002 (−0.062, 0.058)	0.004 (−0.056, 0.063)
Smoking	0.132*** (0.087, 0.177)	0.139*** (0.093, 0.184)
Family History	−0.024 (−0.054, 0.006)	−0.021 (−0.051, 0.010)
Joint Pain	0.674*** (0.662, 0.686)	0.677*** (0.666, 0.689)
Disease Duration	−0.007*** (−0.009, −0.005)	−0.007*** (−0.009, −0.005)
Current Steroid	0.125*** (0.091, 0.159)	0.129*** (0.094, 0.163)
Constant	−0.017 (−0.131, 0.098)	−0.037 (−0.151, 0.077)
Observations	11,282	11,282
R^2	0.728	0.731
Adjusted R^2	0.728	0.731
Note:		$^{*}p < 0.1$;$^{**}p < 0.05$;$^{***}p < 0.01$.

TABLE 7 Comparison between imputation methods after we simulated 10%, 20%, and 30% missing data in the KRRD data set. The RMSE is used to highlight and select the best missing imputation method with the lowest RMSE score.

Method	MAR			MCAR			MNAR		
Missingness rate	10%	20%	30%	10%	20%	30%	10%	20%	30%
Predictive mean matching (PMM)	5.349	6.066	4.944	5.590	4.590	5.135	6.950	7.471	6.516
Random forest (RF)	4.618	5.233	5.234	4.837	4.204	4.539	8.004	7.212	6.737
Classification and regression trees (kNN)	4.107	4.884	4.184	3.820	3.560	3.734	6.236	5.507	5.062
missForest	2.518	3.013	3.032	3.168	2.555	2.871	4.962	4.180	3.631

4 | DISCUSSION

The obtained rheumatoid arthritis (RA) patient data recorded in the Kuwait Registry for Rheumatic Diseases (KRRD) registry were utilized to quantify the Rheumatoid Arthritis Disease Activity Score. All the information was acquired from 1735 patients from public healthcare facilities with permission from the relevant ethical committees. The baseline variables under investigation for every patient included smoking, sex, disease duration, age, nationality, SICCA, RF, ACPA, ANA, family history, treatment class (biologics, cDMARDs), current steroids, comorbidity, DAS28 group, and joint pain.

Systematic errors that existed between the anticipated and noted values were due to the missing value led to outcome bias. However, to get the accurate missing values [45], it is significant to eliminate the bias and apply the optimal approach to guarantee reliability and quality of data analysis. The uncompleted data sets contradicted significantly with the complete data file [46]. The emergence of imputation algorithms has been attributed to their substantial global use.

Imputation methods overcome the existing prejudice in missing values. However, their values may potentially lead to bias in the result. Therefore, they should be used vigilantly. The research utilized numerous variables to define the RA disease activity scores. The application of many factors resulted in data with missing characteristics in various patients, leading to bias results. The focal point was to identify the suitable imputation approach to complete the missing features in the RA data set.

The variation of the missing data ranged from 2% to 66%. At this point, four imputation approaches were assessed, the kNN, PMM, miss Forest, and RF throughout three diverse missing approaches MAR, MNAR, and MCAR with Kuwait Registry for Rheumatic Diseases (KRRD) RA infection data set. Performance evaluations of the imputation methods were done utilizing RMSE values, with the minimum RMSE value showing the best imputation technique.

Multiple imputations (MI) are computationally comprehensive and require estimations. To get enough needed results, several algorithms should be run frequently, where running time increases with more missing data. In our case scenario, MI generated variable approximate almost similar to known approximates than the traditional missing data techniques. MI provided mean and standard error closely similar to the noted values, outperforming the single imputation methods or deletion.

The findings, in this case, are similar to the findings obtained when using a hypothetical data set to differentiate missing data approaches. Putting traditional methods into consideration in all the conducted research, the regression imputation generated the approximate mean, and deletion generated the approximate standard deviation when contrasted to the finalized data sets. The current research was dominant since the variable estimates attained by each

missing data approach could be contrasted to the already known variables of an absolute data set acquired from the clinical setting. The result reveals that is possible to apply missing data methods like MI in the current context [47].

However, despite the effectiveness of the MI method in undertaking the missing data, it is significant to note that the associated problem with missing data cannot be enhanced by any missing data approach. MI and numerous missing data techniques are useful for MAR or MCAR despite their unreliability when data is MNAR. Determination of whether data is MAR or MNAR is often difficult as there is no reliable technique to do so. But, in some clinical or environmental studies (e.g. [48, 4, 49, 14]), either MAR or MCAR are preferable rather than MNAR mechanism.

In general, our results shows that MI using MAR mechanism had the lowest RMSE among the other missingness mechanism (MCAR or MNAR). Compared with complete case analysis, its effectiveness is due to MI's use of information in incomplete cases, while Complete Case Analysis (CCA) is only valid in the case of MAR or MCAR data.[50]. In well-designed studies, such as clinical trials, MAR mechanism is more common than MCAR, because in most cases, observable data explain most of the deficiencies [51].

MI technique sometimes is not the better method even when MCAR or MAR is missing. Concerning the sample size, it is important to note that a small sample size may minimize the accuracy of MI [52]. Additionally, the utilization of MI in longitudinal designs with layered data may present challenges that may need the use of MI algorithms or other approaches other than MI [53, 54, 55]. Another challenge is that statistical packages vary with their ease of usability in respect to the merging variable and test statistics. The provision of numerous missing data methods indicates the benefits of using MI in the clinical surrounding. Additionally, it also indicates the significance of having a comprehensive understanding of the type and the effect of the missing data despite active handling of the data. It is also important to consider factors that may potentially facilitate missing data before the beginning of the research [55]. That way, researchers can measure these factors influencing data missingness and do extensive analysis.

Finally, the variety of data and the negative effects of missing data, and the correlated restraints that come with using traditional approaches to handle the missingness of data are not considered important. The findings show that techniques like MI perform better than the traditional approaches as they facilitate the reintroduction of the difference that would occur upon attaining missing scores. As a result, this reduces bias produced by missing data and enhances the ability to realize meaningful influences. MI and other techniques are fast and elementary to use and their long terms merits are valuable the time taken to learn the techniques and applied within the clinical research setting. From the results, missForest is regarded as the most productive imputation technique with the least RMSE values at 95% credence, reproduced employing 10%, 20%, and 30% missing data. Thereafter, kNN was conducted. The RF and PMM were identified as the least performing imputation techniques. Due to the availability of large data from registered RA patients used, the research and its outcomes are considered robust. Additionally, the imputation method considered and missingness procedures (implemented at 10% to 30% utilizing MAR, MNAR, and MCAR) ameliorated data reliability with notable p-values attained. Also, To check the robustness of the results for the imputation methods particularly when missingness rate is high, we have repeated the new missing imputation using missForest, random forest, kNN and predictive mean matching (PMM) for missingness rate = 30% only and m=25. We found that this is very computationally intensive and the results showed differences in the RMSE scores (m=25 is better than m=5). However, the preferences between imputation methods based on RMSE scores are similar without any changing in ranking order. So, the main results for m=25 has the same conclusion for m=5, which say that MissForest is a highly accurate method of imputation for missing data in KRRD and outperforms other common imputation techniques in terms of imputation error and maintenance of predictive ability with imputed values in clinical predictive models. This approach can be used in registries to improve the accuracy of data, including the ones for rheumatoid arthritis patients.

5 | CONCLUSION

MissForest is a highly accurate method of imputation for missing data in KRRD and outperforms other common imputation techniques in terms of imputation error and maintenance of predictive ability with imputed values in clinical predictive models. This approach can be used in registries to improve the accuracy of data, including the ones for rheumatoid arthritis patients.

acknowledgements

The authors would like to thank KRRD group for their help to provide all the necessary information and data. Also, the author would like to thank all of those who have contributed in data collection and data processing for the registry. Without them, this study will not have been possible.

conflict of interest

Authors have declared that no competing interests exist.

references

[1] Sartori N, Salvan A, Thomaseth K. Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. Computational statistics & data analysis 2005;49(3):937–953.

[2] Branden KV, Verboven S. Robust data imputation. Computational Biology and Chemistry 2009;33(1):7–13.

[3] Frisell T, SP0187 Why Missing Data Is A Problem, and What You shouldn't Do To Solve It. BMJ Publishing Group Ltd; 2016.

[4] Alsaber AR, Pan J, Al-Hurban A. Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018). International Journal of Environmental Research and Public Health 2021;18(3):1333.

[5] Mondelo D, Imputation Strategies for Missing Data in Environment Time Serial for an Unlucky Situation. Springer Berlin Heidelberg; 2006.

[6] Kang H. The prevention and handling of the missing data. Korean journal of anesthesiology 2013;64(5):402.

[7] Stavseth MR, Clausen T, Røislien J. How handling missing data may impact conclusions: a comparison of six different imputation methods for categorical questionnaire data. SAGE Open Medicine 2019;7:2050312118822912.

[8] Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 2004;38(18):2895–2907.

[9] Higgins JP, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. Clinical Trials 2008;5(3):225–239.

[10] Little RJ, Rubin DB. Statistical analysis with missing data, vol. 793. John Wiley & Sons; 2019.

[11] Rubin DB. Statistical analysis with missing data. Wiley; 1987.

[12] Di Zio M, Guarnera U, Luzi O. Imputation through finite Gaussian mixture models. Computational Statistics & Data Analysis 2007;51(11):5305–5316.

[13] Thijs H, Molenberghs G, Michiels B, Verbeke G, Curran D. Strategies to fit pattern-mixture models. Biostatistics 2002;3(2):245–265.

[14] Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clinical epidemiology 2017;9:157.

[15] Van der Heijden GJ, Donders ART, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. Journal of clinical epidemiology 2006;59(10):1102–1109.

[16] Fielding S, Fayers PM, McDonald A, McPherson G, Campbell MK, Group RS, et al. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health and Quality of Life Outcomes 2008;6(1):57.

[17] Moons KG, Donders RA, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. Journal of clinical epidemiology 2006;59(10):1092–1101.

[18] Zhang N, et al. Methodolgical Progress Note: Handling Missing Data in Clinical Research. Journal of hospital medicine 2019;14:E1.

[19] Breiman L. Random forests. Machine learning 2001;45(1):5–32.

[20] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences 2003;43(6):1947–1958.

[21] Bagheri H, Tapak L, Karami M, Amiri B, Cherghi Z. Epidemiological features of human brucellosis in iran (2011-2018) and prediction of brucellosis with data-mining models. Journal of Research in Health Sciences 2019;19(4):e00462.

[22] Amini P, Maroufizadeh S, Hamidi O, Samani RO, Sepidarkish M. Factors associated with macrosomia among singleton live-birth: A comparison between logistic regression, random forest and artificial neural network methods. Epidemiology, Biostatistics and Public Health 2016;13(4).

[23] Fan S, Kind T, Cajka T, Hazen SL, Tang WW, Kaddurah-Daouk R, et al. Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data. Analytical chemistry 2019;91(5):3590–3596.

[24] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–118.

[25] Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology 2014;179(6):764–774.

[26] Liao SG, Lin Y, Kang DD, Chandra D, Bon J, Kaminski N, et al. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC bioinformatics 2014;15(1):1–12.

[27] Stekhoven DJ. missForest: Nonparametric missing value imputation using random forest. Astrophysics Source Code Library 2015;p. ascl–1505.

[28] Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham III CO, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis & Rheumatism 2010;62(9):2569–2581.

[29] Al-Herz A, Al-Awadhi A, Saleh K, Al-Kandari W, Hasan E, Ghanem A, et al. A comparison of rheumatoid arthritis patients in Kuwait with other populations: results from the KRRD registry. Journal of Advances in Medicine and Medical Research 2016;p. 1–11.

[30] Rubin DB, Multiple imputation for survey nonresponse. New York: Wiley; 1987.

[31] Schafer JL. Analysis of incomplete multivariate data. CRC press; 1997.

[32] Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate behavioral research 1998;33(4):545–571.

[33] Van Buuren S. Flexible imputation of missing data. Chapman and Hall/CRC; 2018.

[34] Cutler A, Cutler DR, Stevens JR. Random forests. In: Ensemble machine learning Springer; 2012.p. 157–175.

[35] Oba S, Sato Ma, Takemasa I, Monden M, Matsubara Ki, Ishii S. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 2003;19(16):2088–2096.

[36] Bennett ND, Croke BF, Guariso G, Guillaume JH, Hamilton SH, Jakeman AJ, et al. Characterising performance of environmental models. Environmental Modelling & Software 2013;40:1–20.

[37] Martínez G, Feist E, Martiatu M, Garay H, Torres B. Autoantibodies against a novel citrullinated fibrinogen peptide related to smoking status, disease activity and therapeutic response to methotrexate in cuban patients with early rheumatoid arthritis. Rheumatology International 2020;40:1873–1881.

[38] Choe JY, Bae J, Lee H, Bae SC, Kim SK. Relation of rheumatoid factor and anti-cyclic citrullinated peptide antibody with disease activity in rheumatoid arthritis: cross-sectional study. Rheumatology international 2013;33(9):2373–2379.

[39] Ma JD, Chen CT, Lin JZ, Li QH, Chen LF, Xu YH, et al. Muscle Wasting Aggravates Rheumatoid Arthritis in Elderly Patients as a Mediator 2020;.

[40] Valdiviezo HC, Van Aelst S. Tree-based prediction on incomplete data using imputation or surrogate decisions. Information Sciences 2015;311:163–181.

[41] Junger W, De Leon AP. Imputation of missing data in time series for air pollutants. Atmospheric Environment 2015;102:96–104.

[42] Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC bioinformatics 2019;20(1):1–11.

[43] Tang F, Ishwaran H. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal 2017;10(6):363–377.

[44] Zakaria NA, Noor NM. Imputation methods for filling missing data in urban air pollution data formalaysia. Urbanism Arhitectura Constructii 2018;9(2):159.

[45] Alsaber A, Pan J, Al-Herz A, Alkandary DS, Al-Hurban A, Setiya P, et al. Influence of ambient air pollution on rheumatoid arthritis disease activity score Index. International journal of environmental research and public health 2020;17(2):416.

[46] Forbes D, Hawthorne G, Elliott P, McHugh T, Biddle D, Creamer M, et al. A concise measure of anger in combat-related posttraumatic stress disorder. Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies 2004;17(3):249–256.

[47] Baraldi AN, Enders CK. An introduction to modern missing data analyses. Journal of school psychology 2010;48(1):5–37.

[48] Tsiampalis T, Panagiotakos DB. Missing-data analysis: socio-demographic, clinical and lifestyle determinants of low response rate on self-reported psychological and nutrition related multi-item instruments in the context of the ATTICA epidemiological study. BMC Medical Research Methodology 2020;20:1–13.

[49] Mishra S, Khare D. On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study. J Med Stat Inform 2014;2(1):9.

[50] Little RJ. Regression with missing X's: a review. Journal of the American statistical association 1992;87(420):1227–1237.

[51] Verbeke G. Linear mixed models for longitudinal data. In: Linear mixed models in practice Springer; 1997.p. 63–153.

[52] McKnight PE, McKnight KM, Sidani S, Figueredo AJ. Missing data: A gentle introduction. Guilford Press; 2007.

[53] Enders CK. Applied missing data analysis. Guilford press; 2010.

[54] Enders CK. Analyzing longitudinal data with missing values. Rehabilitation psychology 2011;56(4):267.

[55] Graham JW. Missing data: Analysis and design. Springer Science & Business Media; 2012.