Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic

Shuo Feng
University of Hong Kong School of Public Health
Celestin Hategeka
Harvard University T H Chan School of Public Health
Karen Ann Grépin (kgrepin@hku.hk)
University of Hong Kong Faculty of Medicine: University of Hong Kong Li Ka Shing Faculty of Medicine
https://orcid.org/0000-0003-4368-0045

Research Article

Keywords: Missing Data, Routine Health Information Systems (RHIS), Health Management Information System (HMIS), Health Services Research, Low and middle-income countries (LMICs), Multiple imputation

Posted Date: April 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-422960/v1

Abstract

Background : Poor data quality is limiting the greater use of data sourced from routine health information systems (RHIS), especially in low and middle-income countries. An important part of this issue comes from missing values, where health facilities, for a variety of reasons, miss their reports into the central system. Methods : Using data from the Health Management Information System (HMIS) and the advent of COVID-19 pandemic in the Democratic Republic of the Congo (DRC) as an illustrative case study, we implemented six commonly-used imputation methods using the DRC's HMIS datasets and evaluated their performance through various statistical techniques, i.e., simple linear regression, segmented regression which is widely used in interrupted time series studies, and parametric comparisons through t-tests and non-parametric comparisons through Wilcoxon Rank-Sum tests. We also examined the performance of these six imputation methods under different missing mechanisms and tested their stability to changes in the data. Results : For regression analyses, there was no substantial difference found in the results generated from all methods except mean imputation and exclusion & interpolation when the RHIS dataset contained less than 20% missing values. However, as the missing proportion grew, machine learning methods such as missForest and k -NN started to produce biased estimates, and they were found to be also lack of robustness to minimal changes in data or to consecutive missingness. On the other hand, multiple imputation generated the overall most unbiased estimates and was the most robust to all changes in data. For comparing group means through t-tests, the results from mean imputation and exclusion & interpolation disagreed with the true inference obtained using the complete data, suggesting that these two methods would not only lead to biased regression estimates but also generate unreliable t-test results. Conclusions : We recommend the use of multiple imputation in addressing missing values in RHIS datasets. In cases necessary computing resources are unavailable to multiple imputation, one may consider seasonal decomposition as the next best method. Mean imputation and exclusion & interpolation, however, always produced biased and misleading results in the subsequent analyses, and thus their use in the handling of missing values should be discouraged.

Keywords : Missing Data; Routine Health Information Systems (RHIS); Health Management Information System (HMIS); Health Services Research; Low and middle-income countries (LMICs); Multiple imputation

Introduction

There has been a growing interest in using data sourced from routine health information systems (RHIS) to monitor and evaluate the performance of health programmes, especially in low and middle-income countries (LMICs). Such systems comprise data collected from health facilities at regular time intervals and usually for a pre-defined set of health indicators of interest. Globally, the leading RHIS platform is known as the District Health Information Software 2 (DHIS2), which is currently used in over 72 LMICs (https://www.dhis2.org/) [1]. However, poor data quality is limiting the greater use of data sourced from RHIS in some settings. Missing values are one of the most common and challenging components of poor data quality in RHIS [2] as their presence introduces uncertainty and ambiguity into the data. Also,

missingness often undermines the statistical properties and the performance of estimators developed using incomplete data, thus limiting trust in the results obtained using these data [3].

Strategies to address missing data are not novel topics in the health informatics literature. But while a great number of missing value imputation algorithms have been developed to address this challenge, there are continuing debates concerning which imputation methods are the best in particular scenarios. For example, Waljee et al. [4] have demonstrated the superiority of the local random forest method (e.g., *missForest*) in imputing missing laboratory values, while Hong & Lynn [5] have pointed out that the use of imputed variables from random forest-based approaches could lead to severely biased inference in a simulation study. Studies in other settings suggest that the results generated from multiple imputation are unbiased and could more closely mimic the true data [6, 7]. However, it is generally agreed that no imputation method should be seen uniformly superior in all kinds of datasets [8, 4]. Indeed, the choice of imputation method is highly dependent on the structure of the data. Whether it is from a longitudinal study [9], patients' health records [10], gene expressions [3], or another source will greatly affect the performance of imputation methods and thus inferences drawn from subsequent statistical analyses.

RHIS datasets share common patterns of missing values. Importantly, and unlike other commonly used health datasets, RHIS datasets tend to have missing values primarily on the dependent variable. For example, researchers often want to examine if the number of health facility visits has gone up or down following a public health programme. In this case, it is not uncommon to find that only a proportion of facilities had reported consistently over time, introducing a pattern of missingness that is non-trivial to exclude or impute. Also, given that some facilities are likely to report more consistently than others, simply excluding those facilities with missing values (i.e., listwise deletion) would not only eliminate a large proportion of the sample but also potentially introduce bias into the subsequent statistical estimators. Another approach is to keep all or most facilities but generate imputed data for the time periods with missing values to fill in the holes; however, it is unclear how the missing values should be imputed in RHIS datasets.

A recent systematic review found that researchers are increasingly making use of RHIS data for research and evaluation purposes in LMICs [11]. The study also found that the most commonly used methodology among these research articles was time series analysis to test or account for trends (35%), including 10% involving interrupted time series (ITS). Geostatistical analyses (16%) and pre-post comparisons (15%) were the next popular techniques, while other longitudinal analyses (13%), other cross-sectional analyses (12%), difference-in-difference (7%), and scenario analyses on cost-effectiveness (2%) were also employed. However, this review pointed out that 75% of these research articles had no description of how missing data were managed in their studies. Among the 25% which did address missing values, simply excluding facilities based on certain exclusion criteria was the most common technique, and only a few studies (10%) attempted imputation or other strategies to handle missing values [11].

To date, there has been no evaluation of missing value imputation methods for data sourced from RHIS in LMICs, a gap this article aims to fill. Namely, we evaluate the performance of popular missing value

imputation methods alongside commonly used analysis techniques employed by studies that involve RHIS datasets. Specifically, we use the advent of the COVID-19 pandemic in the Democratic Republic of the Congo (DRC) as an illustrative case study to test the performance of the various imputation methods. Based on our findings, we recommend strategies to handle missing values when using such data in future studies in the DRC and other similar international contexts

Methods

Study Context and Data Source

This study grew out of a broader project that aimed to investigate the impact of infectious disease outbreaks (i.e., Ebola and COVID-19) on the use of health services in the DRC using data from the national health management information system (HMIS), a DHIS2 enabled RHIS [12]. In the DRC's HMIS, health facilities are expected to report the number of visits delivered each month using a standardized paper form. These paper forms are then transferred to the health zone office and are entered into a centralized database by a skilled health professional. Each health facility is normally expected to report on a complete set of health indicators every month. We defined the lack of cases reported for a given indicator at a given health facility for a given month to be a missing value in our study.

This context was selected partially because the DRC represents an interesting and challenging international context in which to test these techniques but also because many members of the research team have been working closely with RHIS data in this setting for years, thus making it a convenient location to undertake this study; however, we believe the DRC's HMIS shares many common features with RHIS in many LMICs, in particular with those in Sub-Saharan Africa. The DHIS2 system began its implementation in the DRC in 2014; however, only in 2017 did it achieve a national-level scale [13]. From the entire sample of 18,138 facilities in the DRC, we identified 5,510 facilities that had reported every month between January 2017 and October 2020. As the COVID-19 pandemic began in the DRC in March 2020, we considered all data before this month as pre-COVID-19 and data collected since March 2020 as after the onset of COVID-19 or during the pandemic.

Imputation Methods

In terms of the data imputation methods evaluated, we first selected the ones utilized in past RHIS studies, i.e., exclusion, interpolation, and mean imputation [11]. We also examined other algorithms that have been used extensively in imputing missing values, namely random forest, multiple imputation, *k* Nearest Neighbour (*k*-NN), and seasonal decomposition (examples by AK Waljee et al., 2013 and DJ Stekhoven & P Bühlmann, 2012) [4,14]. The simplest method, mean imputation, was also included as a baseline for comparison purposes. Table 1 provides a summary of the six imputation methods examined along with a brief technical description for each method. All analyses were conducted through statistical software R [15].

Table 1: Summary of imputation methods

Name of Imputation Method	Description	R package	Level of complexity to implement
1. Mean imputation	Missing values are replaced with the average of the entire non-missing population in the same month.	*N/A*	Easy
2. Exclusion & Interpolation	Firstly, any facilities with three or more consecutive missing monthly reports are excluded. Next, missing values in the remaining facilities are filled with interpolation.	*N/A*	Easy
3. Nonparametric Missing Value Imputation using Random Forest (*missForest*)	*missForest* is a relatively new Random Forest-based method, which treats the variable with missing values as a dependent variable and regresses it against all the other variables in the dataset through a random forest model. This process is repeated iteratively, and in each step, the missing values are filled with a better prediction. The iteration stops when some threshold is met, i.e., when the changes in the imputed values between steps become small enough. This method is popular because of its ability to handle both categorical and numerical data, as well as very little manual parameter tuning required in the implementation [4].	*missForest* (DJ Stekhoven & P Bühlmann, 2012) [14]	Moderate
4. Multiple Imputation	Multiple imputation also treats the variable with missing values as a dependent variable and estimates it based on the rest of the variables. This estimation is repeated multiple times (*M* times) with a random component involved and being slightly different in each estimation to account for the uncertainty in the missing values. *M* datasets with slightly different estimations of the missing values are returned at the end of the estimation procedure and taking an average across the *M* estimations yields an unbiased estimate of the missing values. The multiple imputation by chained equations (*mice*) implementation in R, in particular, enables an iterative estimation of missing values in multiple variables and provides flexibility in imputing both categorical and continuous variables [16].	*mice* (SV Buuren & K Groothuis-Oudshoorn, 2010) [17]	Moderate
5. *k* Nearest Neighbour (*k*-NN)	For each missing data point, the *k*-NN algorithm looks for the other *k* non-missing observations that are the most similar to the missing one, by comparing their distance measures. The missing data is then filled by a weighted average of the *k* neighbouring but non-missing observations, with the weights calculated based on their Euclidean distances to the missing data point. One difficulty in this method is the choice of *k*. In our study, we use the default number of *k*=10 nearest neighbours, but the choice of *k* can be more carefully tuned through cross-validation [18].	*DMwR* (L Torgo, 2010) [19]	Difficult: Users are required to specify the parameter *k*.
6. Seasonal Decomposition	Seasonal decomposition is tailored to the handling of missing values in time series data and can be summarised in three steps. Firstly, it identifies and removes the seasonal component from the original time series. Next, the missing imputation is performed on the deseasonalized series. Finally, the seasonal component is added back to reflect seasonality [20].	*ImputeTS* (S Moritz & T Bartz-Beielstein, 2017) [20]	Easy

In the implementation of methods 3, 4, and 5, we also included leads and lags with one time unit into the imputation, as recommended by [21] that including the time series' own history and future can help predict the time point of interest.

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.

Statistical Analysis

We evaluated the performance of the six imputation techniques mentioned above through the three most commonly used analytical methodologies in RHIS datasets as identified in the systematic review [11]. These methods are:

Simple Linear Regressions;

Segmented Regressions, which is the recommended technique to conduct ITS studies [21] and is widely used in evaluating health system quality improvement interventions when randomization is not possible [22];

Parametric group comparisons through paired t-tests and non-parametric comparisons through paired Wilcoxon Rank-Sum tests, both of which are widely used in pre-post comparison studies.

Missing Data Mechanism

Before the imputation methods can be evaluated, it is important to first understand the missingness mechanism in the dataset. Missingness mechanisms are typically classified as (1) Missing Completely At Random (MCAR), where the probability of being missing is totally random and does not depend on the value of any variables; (2) Missing At Random (MAR), where the missing values in the variable may depend on the known values of other variables in the data but not on the missing variable itself; and (3) Missing Not At Random (MNAR), where the missingness of a variable could depend on the missing variable itself [24].

If data are believed to MNAR, it is generally recommended to improve the data quality by re-collecting data rather than using an imputation method because the missing pattern is not observed in the dataset [10,25]. On the other hand, if the data are believed to be MCAR, i.e., the probability of a data point being missing is totally random and independent from any of the other variables, then a complete case analysis in which missing values are simply removed would generate unbiased results in subsequent statistical analyses [26]. If, however, the data are believed to be MAR, i.e., the missing pattern can be fully identified using the observed data, some algorithms can be applied to impute the missing values, resulting in a new complete dataset with imputed values. This new complete dataset can then be used to conduct further analyses.

To simulate a scenario where the RHIS data were missing at random, we inserted missing values into an HMIS dataset consisting of 5,510 always-reporting facilities from the DRC's HMIS as follows: the monthly total number of clinical visits at time and for facility was set to missing depending on the facility 's location (city and province), facility type (one of Hospital, Health Post, or Health Centre), time (the number of months elapsed since January 2017), season (a four-level categorical variable: 1 for January to March, 2 for April to June, 3 for July to September, and 4 for October to December), log population, and a binary indicator of the COVID-19 pandemic (0 for January 2017 through February 2020, and 1 otherwise), through the following equation:

$$logit\left(\mathbb{P}\left(visits_{ij} = missing\right)\right) = \beta_o + \beta_1 time_i + \beta_2 season_i + \beta_3 COVID_i + \beta_4 FacilityType_j + \beta_5 Province_j + \beta_6 City_j + \beta_7 log\left(Pop_j\right)$$

With this formula, we produced six datasets with 5%, 10%, 15%, 20%, 25%, or 30% of the monthly visits set to missing, respectively. Next, each of the six missing value imputation methods described above was used to impute missing values in each dataset, with imputation bias and Root Mean Square Error (RMSE) calculated to compare the imputed

values with the true observed values, where $bias = \frac{\sum_{i=1}^{n}\left(x_i^{imputed} - x_i^{true}\right)}{n}$ and $RMSE = \frac{\sqrt{\sum_{i=1}^{n}\left(x_i^{imputed} - x_i^{true}\right)^2}}{n}$.

Consecutive Missingness

Occasionally, facilities may consecutively miss making their monthly reports rather than following a pattern that would instead be considered MAR. Figure 1 summarizes the number of facilities with no missing monthly reports, with missing reports but no consecutive missing reports, exactly two consecutive missing reports, exactly three consecutive missing reports, and at least four consecutive missing reports, respectively. We observed that there was a considerable number of facilities with at least four consecutive missing reports (4,446 out of 18,138 facilities, or approximately 25%), which led us to also consider the performance of imputation methods in datasets with consecutive missing values. Specifically, we generated two additional datasets with 15% and 30% consecutive facility-month reports randomly set to missing.

Subsequent Analyses

After the imputed datasets were constructed, we then performed three types of analyses on each, i.e., simple linear regression, segmented regression analysis of ITS, and parametric and non-parametric tests for comparing groups. For simple linear regression, the facility-level monthly total number of clinical visits was used as the target variable and the incidence rate ratio (IRR) was estimated, with the following explanatory variables:

Time – a discrete variable counting the number of months elapsed since January 2017;

COVID – a binary variable indicating the presence of COVID-19 pandemic, i.e., 0 for January 2017 through February 2020, and 1 otherwise;

Log Population – a continuous variable capturing the log-transformed population size of the health zone where the facility is located;

Facility type – a categorical variable specifying the type of facility. Possible values are Hospital, Health Post, or Health Centre;

Province – a categorical variable was added to account for the 26 provinces in the DRC;

Season – a categorical variable to control for each of the 4 seasons as there are known to be seasonal variations in the use of health services in comparable settings [27].

For segmented regressions, we considered a facility-level mixed-effect segmented Poisson regression model, with the same target and explanatory variables as in the simple linear regression described above.

For pre-post comparisons, we conducted both parametric t-tests and non-parametric Wilcoxon Rank-Sum tests on each of the imputed datasets to examine if there were statistically significant differences in the mean number of clinical visits before vs. during the COVID-19 pandemic using paired t-tests and to examine if there was a location shift in the median number of clinical visits before vs. during the pandemic using paired Wilcoxon Rank-Sum tests. The number of monthly clinical visits was selected as a good overall measure of health services utilization in DRC, and this context was chosen because a decrease in the use of health service was found in Kinshasa, DRC following the onset of the COVID-19 pandemic [28].

Stability

RHIS databases are typically updated regularly. For example, in the DRC, there is a monthly update of the datasets that reflects the accretion of reports obtained from health facilities, including some that may have been submitted with a delay. These RHIS datasets are meant to be updated frequently, hence it is important to ensure consistency in the imputed values as well as the subsequent estimators obtained using these data from month to month. We therefore tested each imputation method's stability to minimal changes in the dataset (i.e., with only two months of data removed). In particular, we designed a scenario where the last two months (September and October 2020) were removed as well as a

scenario where two random months of data (i.e., two months chosen randomly from the entire dataset of 46 months with a ten-fold cross-validation) were removed. We compared the performance of each imputation method on the datasets generated under those two scenarios with its performance on the original dataset to evaluate the method's stability. Besides that, we also repeated each imputation model on the original dataset but with another random starting point, as many machine learning optimization algorithms are found to be starting-point dependent [29], and thus the choice of starting point could potentially have an impact on the convergence and performance of the imputation method.

Results

Levels of Missing Data in the DRC's HMIS Dataset

To provide context for this study, we first calculated the actual percentage of missing data in the DRC's HMIS. In the original HMIS dataset, we observed a higher report missing rate in health posts relative to health centres or hospitals (Fig. 2). However, the percentage of missing data has greatly decreased over time for all types of facilities. By 2020, there were only approximately 20% missing monthly reports in health posts and about 5% missing in health centres and hospitals.

In addition to the total number of clinical visits, we also examined the levels of missing data for several other essential health services, including visits for common infectious diseases (uncomplicated pneumonia, uncomplicated diarrhea, uncomplicated malaria, and rapid diagnostic tests for malaria), visits for maternal health services (antenatal consultations, institutional deliveries, and postnatal consultations), new diagnoses of non-communicable diseases (diabetes and hypertension), and vaccinations (DTP, BCG, OPV, and PVC-13). More details about the levels of missingness for these indicators can be found in Figures A.1 to A.4b in the Supplementary Materials.

Bias and RMSE

Firstly, we were interested in examining how much the imputed values themselves deviated from the true complete data in terms of bias (Fig. 3) and RMSE (Fig. 4), with different proportions of missing values inserted.

In our study, we observed the bias and RMSE to grow as more missing values were introduced. With the same level of missing data, we found no material difference in the bias among the imputed values from different imputation methods, except *missForest* which has erroneously produced very biased values when the missing proportion was relatively large (i.e., when more than 20% missing values present).

In terms of RMSE, the data generated using mean imputation always had the largest RMSE, while the RMSE for the data generated using the exclusion & interpolation method was the most consistent across

different missing proportions and was the smallest when the missing proportion was large. The other four imputation methods, however, were found not to be much different from each other regarding their RMSEs.

Estimated Coefficients Under MAR Assumption

Figures 5a and 5b show the estimated coefficients for the discrete variable *time*, which counts the number of months elapsed since January 2017, and for the binary variable *COVID*, which equals to 0 for January 2017 through February 2020 and equals to 1 otherwise, both from the fitted simple linear regressions. Their corresponding 95% confidence intervals were shown as error bars in the graphs. The estimated coefficients and confidence intervals for the other variables exhibited very similar behaviours, which are provided in Figures A.6a through A.6c in the Supplementary Materials. Figures 6a and 6b show the estimated level and trend change IRRs with their 95% confidence intervals from segmented regressions of ITS.

From both sets of graphs, we observed that regardless of the subsequent analysis conducted and regardless of the variable type, the performance of all imputation methods – except mean imputation and exclusion & interpolation – was similar when there were only 5% or 10% missing values present in the data. As the missing proportions grew, however, the performance, i.e., the estimated coefficient and confidence intervals, started to deviate from the true estimates generated using the complete HMIS data. This separation among the estimates was found to be continuously aggravated as more missing data were inserted. On the other hand, the estimates based on the datasets generated using mean imputation or exclusion & imputation deviated from the true estimates immediately after even a very small number of missing values had been introduced.

Pre-Post Comparisons

As the next most popular techniques performed in RHIS datasets, paired t-tests and paired Wilcoxon Rank-Sum tests were conducted to compare the numbers of monthly facility-level clinical visits from January 2019 to October 2019 versus the numbers from January 2020 to October 2020, with the resultant p-values provided in Tables 2a and 2b.

As shown in Table 2a (i.e., paired t-tests), only the datasets generated using mean imputation or exclusion & interpolation have produced a different inference result (i.e., a p-value great than 0.05) versus using the true complete data (i.e., a p-value substantially less than 0.05). For Table 3b (i.e., paired Wilcoxon Rank-Sum tests), however, all the estimated p-values were found to be less than 0.01.

Table 2
a: p-values obtained from comparing group means using paired t-tests

Imputation Method	5% Missing	10% Missing	15% Missing	20% Missing	25% Missing	30% Missing
Complete Data	0.008 †	0.008 †	0.008 †	0.008 †	0.008 †	0.008 †
Mean	0.044 *	0.080	0.084	0.101	0.187	0.598
Exclusion & Interpolation	0.021 *	0.096	0.118	0.162	0.103	0.058
Random Forest	0.005 †	0.001 †	0.003 †	0.007 †	0.001 †	0.013 *
Multiple Imputation	0.006 †	0.005 †	0.013 *	0.002 †	0.003 †	0.018 *
k-NN	0.006 †	0.008 †	0.009 †	0.005 †	< 0.001 †	< 0.001 †
Seasonal Decomposition	0.001 †	0.001 †	< 0.001 †	< 0.001 †	< 0.001 †	< 0.001 †

Table 2
b: p-values obtained from comparing group medians paired Wilcoxon Rank-Sum tests

Imputation Method	5% Missing	10% Missing	15% Missing	20% Missing	25% Missing	30% Missing
Complete Data	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16
Mean	4.00E-14	5.71E-12	1.40E-07	4.59E-05	2.84E-04	6.13E-04
Exclusion & Interpolation	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	< 2.2E-16	5.21E-15
Random Forest	< 2.2E-16	5.60E-16	3.28E-15	3.99E-13	1.36E-05	8.09E-10
Multiple Imputation	< 2.2E-16	< 2.2E-16	< 2.2E-16	5.36E-11	5.76E-09	4.70E-12
k-NN	< 2.2E-16	< 2.2E-16	2.15E-15	4.70E-14	7.23E-13	1.08E-08
Seasonal Decomposition	2.51E-15	2.51E-15	3.74E-15	3.84E-11	1.10E-10	3.33E-10

† for p-value < 0.01, * for p-value < 0.05, and no annotation for p-value greater than 0.05

Consecutive Missing and Stability

Figures 7 and 8 plot the estimated coefficients for the variables *time* from simple linear regression and the estimated level change IRRs from segmented regressions with their corresponding 95% confidence intervals, on the datasets with 15% and 30% missing values inserted in different ways. Columns from left to right are estimates on (a) original dataset with missing values inserted under the MAR assumption, (b) missing values inserted consecutively, (c) stability: dataset with last two months removed, (d) stability: dataset with two random months removed and with 10-fold cross-validation, and (e) stability: original

dataset with missing values inserted under the MAR assumption but the imputation algorithms started with a different starting point.

As can be seen from Figs. 7 and 8, the datasets imputed using seasonal decomposition and the two machine learning algorithms, i.e., *missForest* and *k*-NN, were found to be the most vulnerable to changes in the data, such as missing values inserted differently (consecutive missing), or minimal changes in the data (two months data points removed). The other imputation methods were relatively stable for minimal changes in the data. Also, whether the missing values were inserted under MAR or consecutive missing had little impact on all imputation methods other than *missForest* and *k*-NN. The behaviour for the other variables was almost identical to the two shown. They are provided in Figures A.7a through A.7e in the Supplementary Materials.

Discussion

In this study, we observed no substantial difference in the bias produced by all imputation methods when the missing proportion was relatively small, i.e., when less than 20% missing values present in the data (Fig. 3). However, as the missing percentage grew, datasets imputed using *missForest* quickly became the most biased. This finding is consistent with previous evidence that implementing individual tree estimation through *missForest* could systematically lead to biased estimates, especially for non-normal, skewed data such as the count data that we had in our study [30]. Further, we found that the bias in imputed values became more aggravated as more missing values were introduced into the data. On the other hand, *k*-NN has consistently generated relatively unbiased data, even when the proportion of missing was large. However, although the values themselves generated using *k*-NN were the least biased, the estimators constructed using such an imputed dataset were not guaranteed to be unbiased. In fact, our study suggests that *k*-NN could lead to a more biased and unstable segmented regression estimates than any other methods when the missing proportion was large.

For RMSE, although the data generated using mean imputation always had the largest RMSE (Fig. 4), the exclusion & interpolation approach outperformed all other imputation methods when the missing proportion was sufficiently large. But this merit could be attributable to the fact that this method had already excluded those facilities suffering the most from missing values, and its RMSE was calculated solely based on the subpopulation of the most consistently-reporting facilities. On the other hand, the data imputed using seasonal decomposition produced the least RMSE when there were few missing values and outperformed the other imputation methods, except exclusion & interpolation, as the missing percentage grew. We otherwise observed no material difference between the other three methods.

In terms of estimated coefficients and confidence intervals from simple linear regressions, as more missing values were introduced into the data, there was a growing deviation between the regression coefficient estimates and the true estimates based on the complete data (Figs. 5a and 5b). Specifically, with only 5%, 10%, or 15% missing inserted, all imputation methods, except mean imputation and exclusion & interpolation, performed reasonably well in estimating the regression coefficients and their

confidence intervals. However, as more missing values were inserted, most imputation methods started to produce erroneous coefficient estimates that deviated from the true estimates using the complete data.

Similar to the case with simple linear regression, all imputation methods, except mean imputation and exclusion & interpolation, generated accurate level and trend change estimates for segmented regressions when the missing proportion was relatively small (Figs. 6a and 6b). As the missing proportion grew, estimates from the datasets imputed using *missForest*, multiple imputation, and seasonal decomposition still maintained a good level of accuracy, while *k*-NN has produced severely biased estimates, especially in estimating the level change IRRs. Multiple imputation has generated the overall most unbiased estimates for both level and trend change IRRs, aligning with the findings from WR Myers using clinical trial data [7].

For pre-post comparison, even though all datasets produced a p-value less than 0.01 in the paired Wilcoxon Rank-Sum tests (Table 1b), we did observe some notable discrepancies in the results from paired t-tests (Table 1a). Specifically, while the original complete dataset demonstrated sufficient statistical evidence to reject the null hypothesis of no difference in the monthly number of clinical visits between pre- and during-COVID-19 at a 95% significance level (i.e., a p-value much less than 0.05), datasets imputed using mean imputation or exclusion & interpolation however produced a p-value greater than 0.05 when at least 10% missing values were inserted. This suggests that the use of simple mean imputation or exclusion & interpolation would not only lead to biased regression estimates but could also generate unreliable t-test results. All the other imputation methods, however, generated consistent conclusions (i.e., a p-value less than 0.05) with the complete data, regardless of the amount of missingness introduced.

In terms of stability, results from the datasets imputed using seasonal decomposition or machine learning algorithms (i.e., *missForest* and *k*-NN) were the most sensitive to any changes in the data or a different starting point, especially when the missing proportion was large (Figs. 7 and 8). This lack of stability makes them less ideal to use when the missing proportion is large. On the other hand, multiple imputation was found to be the most robust to all changes in the data. We also verified that, whether the missing values were inserted under the MAR assumption or inserted consecutively had very little impact on all imputation methods aside from *missForest* and *k*-NN, both of which performed relatively poorly.

In practice, exclusion & interpolation is the most widely used method to deal with missing values among public health researchers who work closely with RHIS data [11]. Our study, however, suggests that the data imputed using this method can potentially lead to severely biased estimates and thus incorrect inference. Instead, the other statistically reliable but simple-to-implement imputation methods, especially the use of multiple imputation, should be encouraged. These methods can be implemented easily through existing packages from various statistical software, including the freely-available software – R [15].

Though our study has shown multiple imputation to outperform all other imputation methods because of its consistency across varying levels of missingness and also its robustness to all types of changes in

the data, it may be challenging to implement multiple imputation in RHIS datasets due to its lengthy computing time. RHIS datasets typically cover a considerable number of facilities and the number of observations is multiplied by the number of time periods for which the reports are collected. This massive amount of data, as well as the nature of multiple imputation to repeat the imputation algorithm several times, increases the computing time exponentially. This issue may be particularly challenging under LMIC settings where more limited computing resources may be available. In the scenario where the computing time for multiple imputation is unavailable, we recommend the use of seasonal decomposition as the next best method to use, although the researchers must be cautious when the missing proportions are large as this method is observed to be relatively unstable to the changes in the data when there are approximately 30% missing values present.

We believe our study can be well generalized to RHIS in other LMICs. Firstly, the DRC is a low-income country with a relatively weak and dysfunctional health system [31]. In particular, the inefficiencies in the DRC's health systems, including the failure to provide complete and consistent facility-level reports and human errors introduced by handling and transferring all the paper reports manually, are likely to be observed in many other LMICs. Also, among all the LMICs, the DRC is experiencing one of the most severe data quality issues as we illustrated in our study. The DRC's HMIS has relatively high levels of missingness among LMICs whose health systems had been evaluated for their completeness in the literature. For example, Rwanda has a similar national HMIS but has been switched from paper forms to electronic HMIS since 2008. By 2012, there was no more than 5% missing reports for key indicators including general clinical visits, maternal health services, and vaccinations in its HMIS [32].

This study has several limitations. Firstly, we only included in our study those facilities that had reported every month between January 2017 and October 2020, and this study population (5,510 facilities) is a small subset of the entire population (18,138 facilities) in DRC. The entire population could also be further examined to confirm the generalizability of our findings. It would also be interesting to valid externally that whether our results generalize to the RHIS data in another country. Also, we only considered regression, segmented regressions of ITS, and pre-post comparison in our analysis. While these three methodologies are among the most commonly used methods to analyse RHIS data [11], it is important to note the growing prevalence of machine learning techniques in the use of health data. Future research could also examine whether machine learning or deep learning algorithms developed based on RHIS datasets imputed by different missing value imputation methods could lead to similar conclusions.

Conclusion

As Brock et al. (2008) [8] have pointed out, no imputation method should be seen uniformly superior in all kinds of datasets. Consistent with this message, our study finds that the performance of imputation methods based on RHIS data varies from that under other data contexts. Specifically, when the missing proportion was relatively low (i.e., less than 20%), we did not observe any substantial differences in the results generated from the four imputation methods evaluated (all except mean imputation and exclusion

& interpolation), while seasonal decomposition slightly outperformed the other ones due to a lower RMSE. As the missing proportion grew larger (i.e., when at least 20% missing values present), *missForest* and *k*-NN started to produce biased estimates, and they were also found to be lack of robustness to minimal changes in the data or consecutive missingness.

In terms of pre-post comparisons, while we do not encourage the use of such direct comparisons, we found that the true complete dataset as well as the datasets imputed using *missForest*, multiple imputation, *k*-NN, and seasonal decomposition all demonstrated sufficient statistical evidence to reject the null hypothesis of no difference in the monthly number of clinical visits between pre- and during-COVID-19 at a 95% significance level. The results from mean imputation and exclusion & interpolation, however, disagreed with what we found using the complete data, suggesting that these two methods would not only lead to biased regression estimates but could also generate unreliable t-test results.

In conclusion, when the missing proportion was relatively small, all methods except mean imputation and exclusion & interpolation produced very similar results and performed well. Therefore, it does not make a substantial difference which of the four imputation methods to choose when levels of missingness are low. However, with at least 20% missing values present, *missForest* and *k*-NN started to provide biased estimates, while multiple imputation was found to be not only the most consistent to the increasing amount of missingness introduced but also the most stable to all types of changes in the data. Overall, we therefore recommend the use of multiple imputation in addressing missing values in RHIS datasets. However, in cases necessary computing resources are unavailable to multiple imputation, one may consider seasonal decomposition as the next best method to use. Mean imputation and exclusion & interpolation, on the other hand, always produced the most biased and misleading results in the subsequent simple linear regression, segmented regressions of ITS, and pre-post comparison tests, and thus their use in handling missing values for RHIS data should be discouraged.

List Of Abbreviations

Abbreviation	Explanation
RHIS	Routine Health Information Systems
HMIS	Health Management Information System
DHIS2	District Health Information Software 2
DRC	Democratic Republic of the Congo
LMICs	Low and Middle-Income Countries
ITS	Interrupted Time Series
missForest	Nonparametric Missing Value Imputation using Random Forest
mice	Multiple Imputation by Chained Equations
k-NN	*k*-Nearest Neighbour
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
RMSE	Root Mean Square Error

Declarations

Ethics approval and consent to participate

We used a research protocol that had been approved by the Ethics Committees at Wilfrid Laurier University (Canada) and Kinshasa School of Public Health (DRC). We also obtained authorization from the Ministry of Public Health to utilise these data to evaluate the impact of the pandemic on health service utilization.

Consent for publication

Not applicable

Availability of data and materials

We received authorization from the Ministry of Public Health to use these data to evaluate the impact of the pandemic on health service utilization. However, the dataset is not publicly available and researchers

who wish to use these data are required to also obtain authorization from the Ministry of Public Health at the DRC. (edited)

Competing interests

S.F., C.H. and K.G. declare no competing interests.

Funding

International Development Research Centre, Rapid Research Fund for Ebola Virus Disease Outbreak, Grant #108966-002 (K.G. PI).

Authors' contributions

S.F. and K.G. conceived of this work and wrote the manuscript. S.F. also contributed to the data analysis in this paper. S.F., C.H., and K.G. all contributed to the study design and reviewing the manuscript. All authors have approved the final version of this manuscript and take accountability for all aspects of the manuscript.

Acknowledgements

Not applicable

References

1. DHIS in Action [Internet]. dhis2. Available from: https://dhis2.org/

2. Hoxha K, Hung YW, Irwin BR, Grépin KA. Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: A systematic review. HIM J. 2020 Jun 30;183335832092872.

3. Schmitt P, Mandel J, Guedj M. A Comparison of Six Methods for Missing Data Imputation. J Biom Biostat [Internet]. 2015 [cited 2021 Apr 9];06(01). Available from: https://www.omicsonline.org/open-

access/a-comparison-of-six-methods-for-missing-data-imputation-2155-6180-1000224.php?aid=54590

4. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open [Internet]. 2013;3(8). Available from: https://bmjopen.bmj.com/content/3/8/e002847

5. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Medical Research Methodology. 2020 Jul 25;20(1):199.

6. Christina M, Zhaohui S, Daniel W. Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [Internet] [Internet]. PubMed. 2018 [cited 2021 Apr 9]. Available from: https://pubmed.ncbi.nlm.nih.gov/29671990/

7. Myers WR. Handling Missing Data in Clinical Trials: An Overview. Drug Information Journal. 2000 Apr 1;34(2):525–33.

8. Brock GN, Shaffer JR, Blakesley RE, Lotz MJ, Tseng GC. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics. 2008 Jan 10;9(1):12.

9. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol. 2018 Dec;18(1):168.

10. Penny KI, Atkinson I. Approaches for dealing with missing data in health care studies. Journal of Clinical Nursing. 2012 Oct 1;21(19pt20):2722–9.

11. Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Serv Res. 2020 Dec;20(1):790.

12. Hung YW, Law MR, Cheng L, Abramowitz S, Alcayna-Stevens L, Lurton G, et al. Impact of a free care policy on the utilisation of health services during an Ebola outbreak in the Democratic Republic of Congo: an interrupted time-series analysis. BMJ Global Health. 2020 Jul 1;5(7):e002119.

13. Mapping the Stages of MEASURE Evaluation's Data Use Continuum to DHIS 2: An Example from the Democratic Republic of the Congo | Evaluate [Internet]. The Mesaure Evaluation Blog. 2019 [cited 2021 Apr 9]. Available from: https://measureevaluation.wordpress.com/2019/06/06/mapping-the-stages-of-measure-evaluations-data-use-continuum-to-dhis-2-an-example-from-the-democratic-republic-of-the-congo/

14. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112–8.

15. The R Foundation. R: The R Project for Statistical Computing [Internet]. [cited 2021 Apr 9]. Available from: https://www.r-project.org/

16. Wulff J, Ejlskov L. Multiple Imputation by Chained Equations in Praxis: Guidelines and Review. Electronic Journal of Business Research Methods. 2017 Apr 1;15:2017–58.

17. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011 Dec 12;45(1):1–67.

18. Varmuza K, Filzmoser P, Hilchenbach M, Krüger H, Silén J. KNN classification — evaluated by repeated double cross validation: Recognition of minerals relevant for comet dust. Chemometrics and Intelligent Laboratory Systems. 2014 Nov 15;138:64–71.

19. Torgo L. Data mining with R: learning with case studies. Boca Raton: Chapman & Hall/CRC; 2011. 289 p. (Chapman & Hall/CRC data mining and knowledge discovery series).

20. Moritz S, Bartz-Beielstein T. imputeTS: Time Series Missing Value Imputation in R. The R Journal. 2017;9(1):207.

21. Honaker J, King G. What to Do about Missing Values in Time-Series Cross-Section Data. American Journal of Political Science. 2010;54(2):561–81.

21. Taljaard M, McKenzie JE, Ramsay CR, Grimshaw JM. The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care. Implementation Sci. 2014 Dec;9(1):77.

22. Hategeka C, Ruton H, Karamouzian M, Lynd LD, Law MR. Use of interrupted time series methods in the evaluation of health system quality improvement interventions: a methodological systematic review. BMJ Global Health. 2020 Oct 1;5(10):e003567.

23. Li L, Li Y, Li Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transportation Research Part C: Emerging Technologies. 2013 Sep 1;34:108–20.

24. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2019 Apr 23.

25. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health. 2004;25:99–117.

26. Humphries M. Missing Data & How to Deal: An overview of missing data [Internet]. Available from: https://liberalarts.utexas.edu/prc/_files/cs/Missing-Data.pdf

27. Ataguba JE. Socio-economic inequality in health service utilisation: Does accounting for seasonality in health-seeking behaviour matter? Health Econ. 2019 Nov;28(11):1370–6.

28. Hategeka C, Carter SE, Chenge FM, Katanga EN, Lurton G, Mayaka SM-N, et al. Impact of the COVID-19 pandemic and response on the utilisation of health services during the first wave in Kinshasa, the Democratic Republic of the Congo. medRxiv. 2021 Apr 10;2021.04.08.21255096.

29. Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems [Internet]. Berlin, Heidelberg: Springer; 2000. p. 1–15. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/3-540-45014-9_1#citeas

30. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Medical Research Methodology. 2020 Jul 25;20(1):199.

31. World Health Organization. Improving health system efficiency: Democratic Republic of the Congo: improving aid coordination in the health sector [Internet]. Geneva: World Health Organization; 2015. Available from: https://apps.who.int/iris/handle/10665/186673

32. Nisingizwe MP, Iyer HS, Gashayija M, Hirschhorn LR, Amoroso C, Wilson R, et al. Toward utilization of data for program management and evaluation: quality assessment of five years of health management information system data in Rwanda. Global Health Action. 2014 Dec;7(1):25829.

Figures

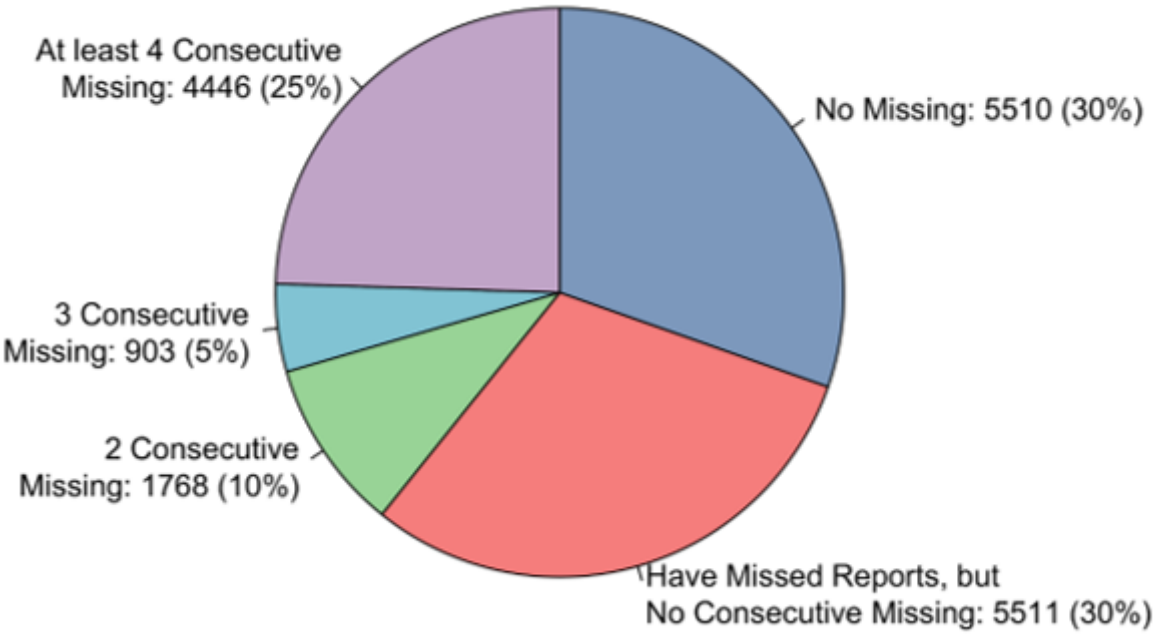

Figure 1

Summary of missing reports

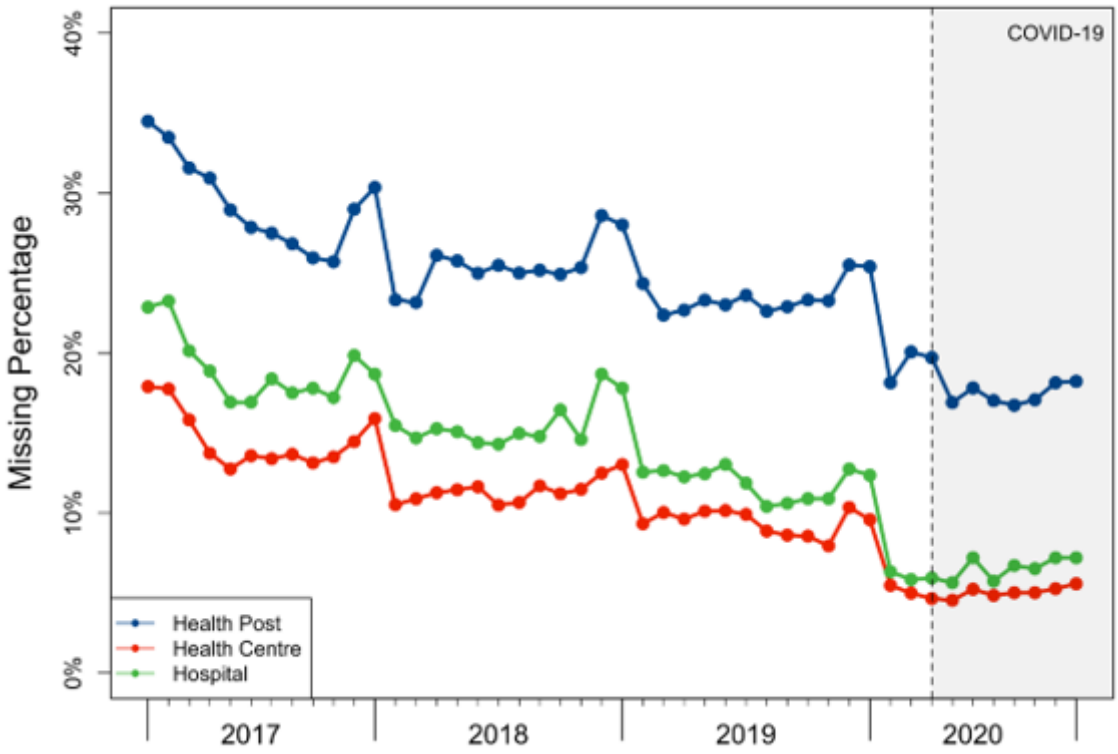

Figure 2

Missing percentages for monthly total visits by facility level

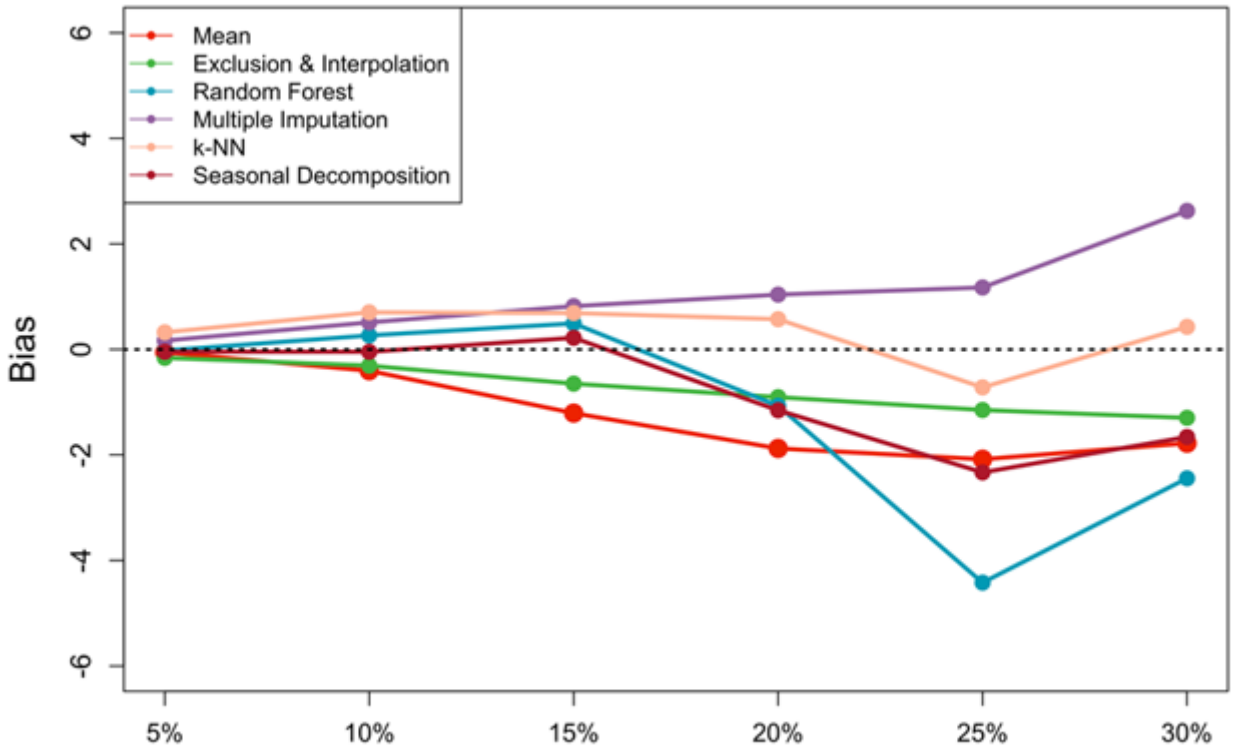

Figure 3

Bias between the imputed and true data with different missing percentages

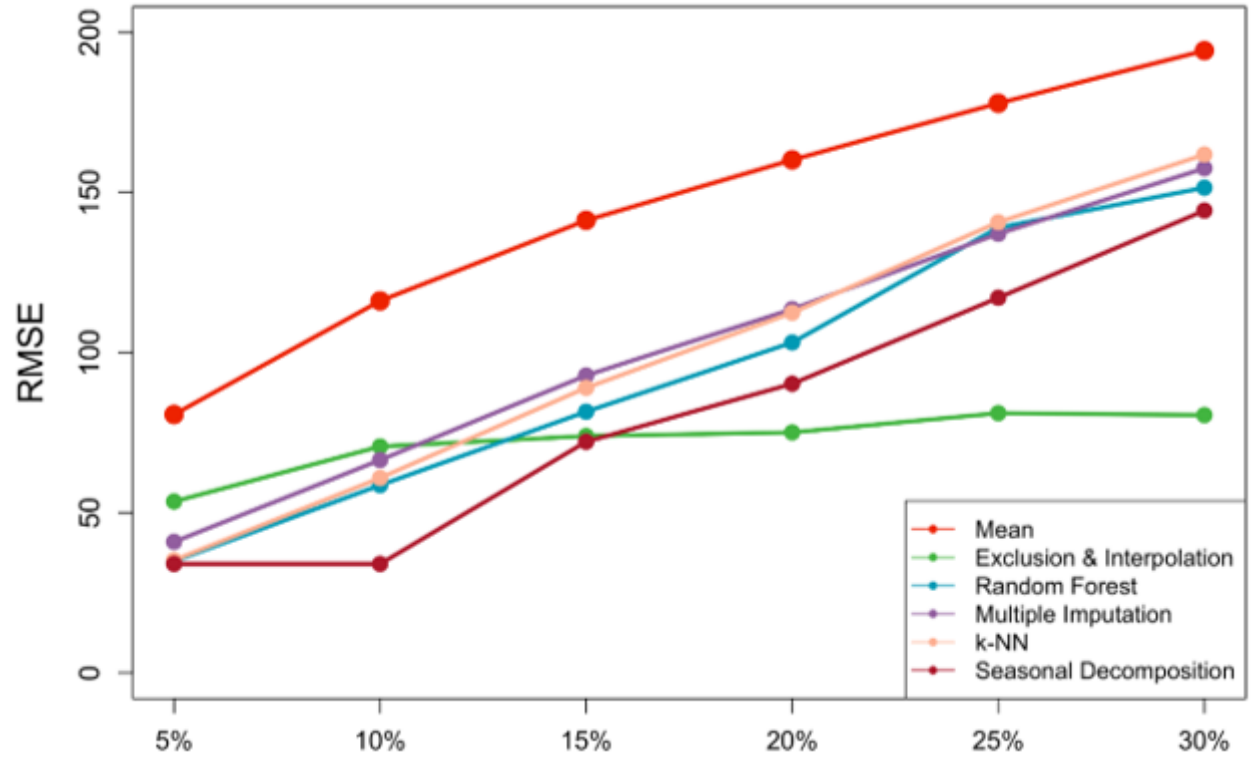

Figure 4

RMSE between the imputed and true data with different missing percentages

Figure 5

a: Estimated coefficients with 95% confidence intervals for the variable time from simple linear regression. b: Estimated coefficients with 95% confidence intervals for the variable COVID from simple linear regression

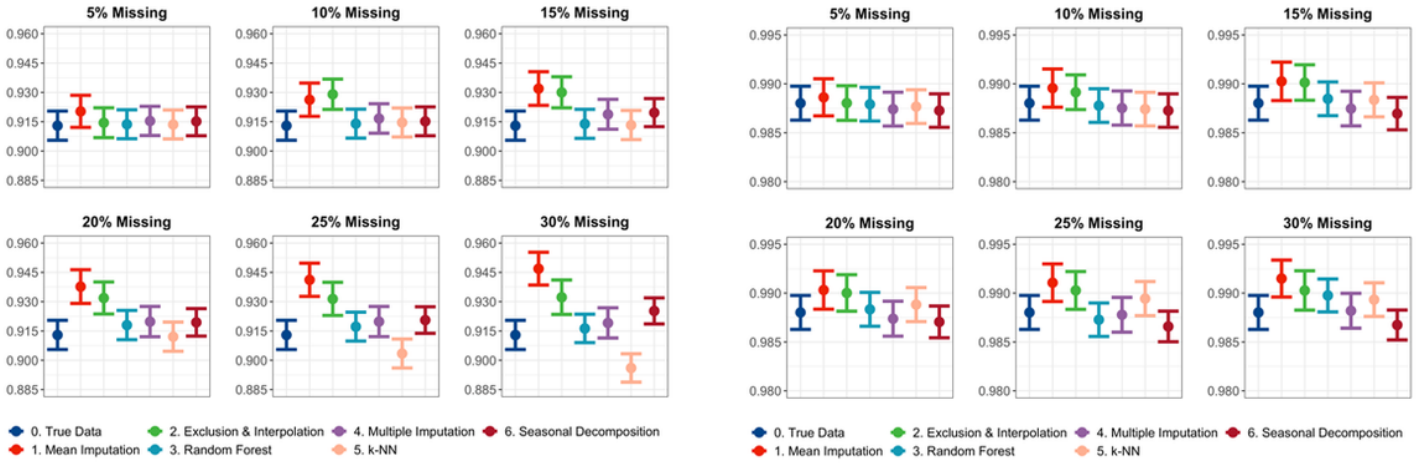

Figure 6

a: Estimated coefficients with 95% confidence intervals for the level change IRR from segmented regressions. b: Estimated coefficients with 95% confidence intervals for the trend change IRR from segmented regressions

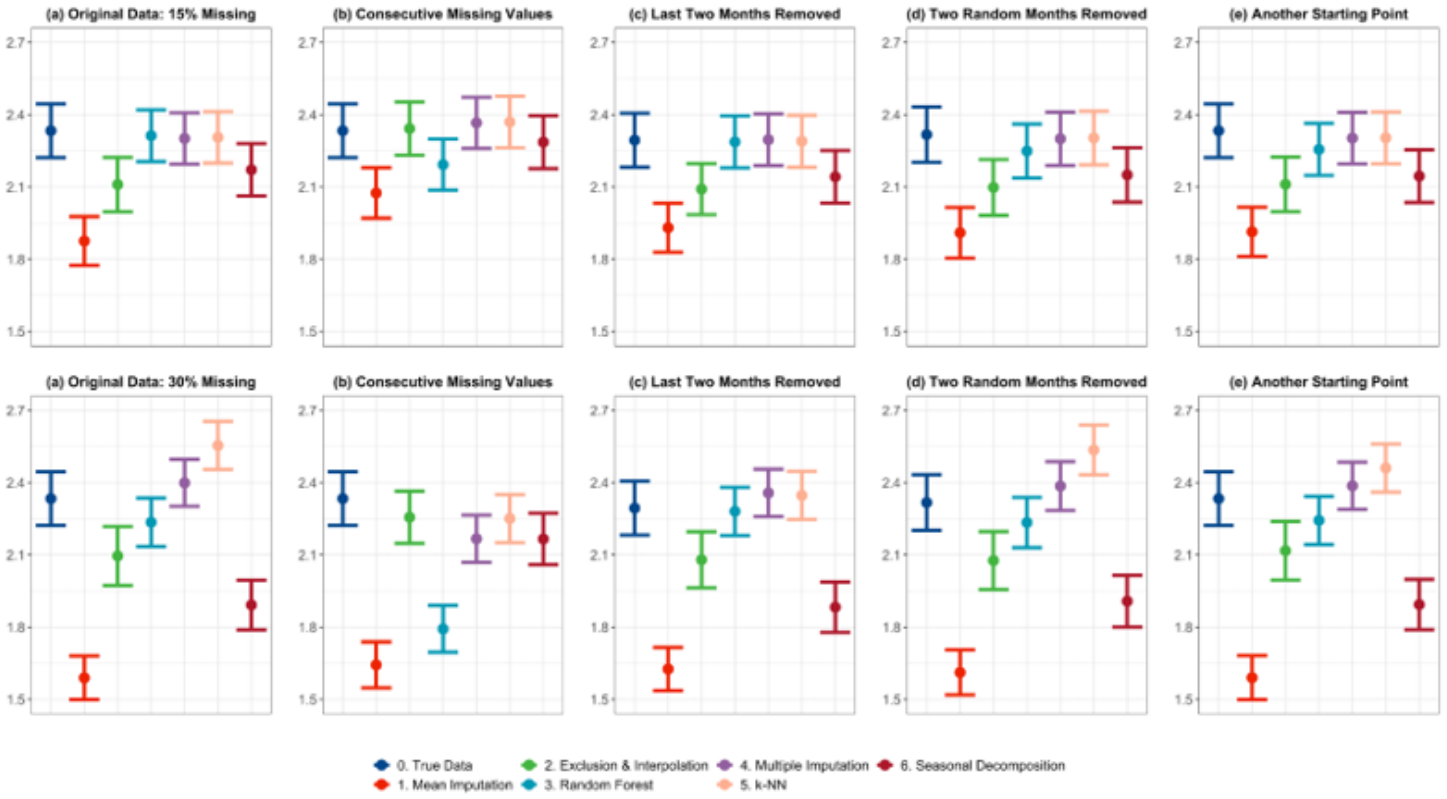

Figure 7

Estimated coefficients and 95% C.I.s for variable time with missing values inserted under different scenarios

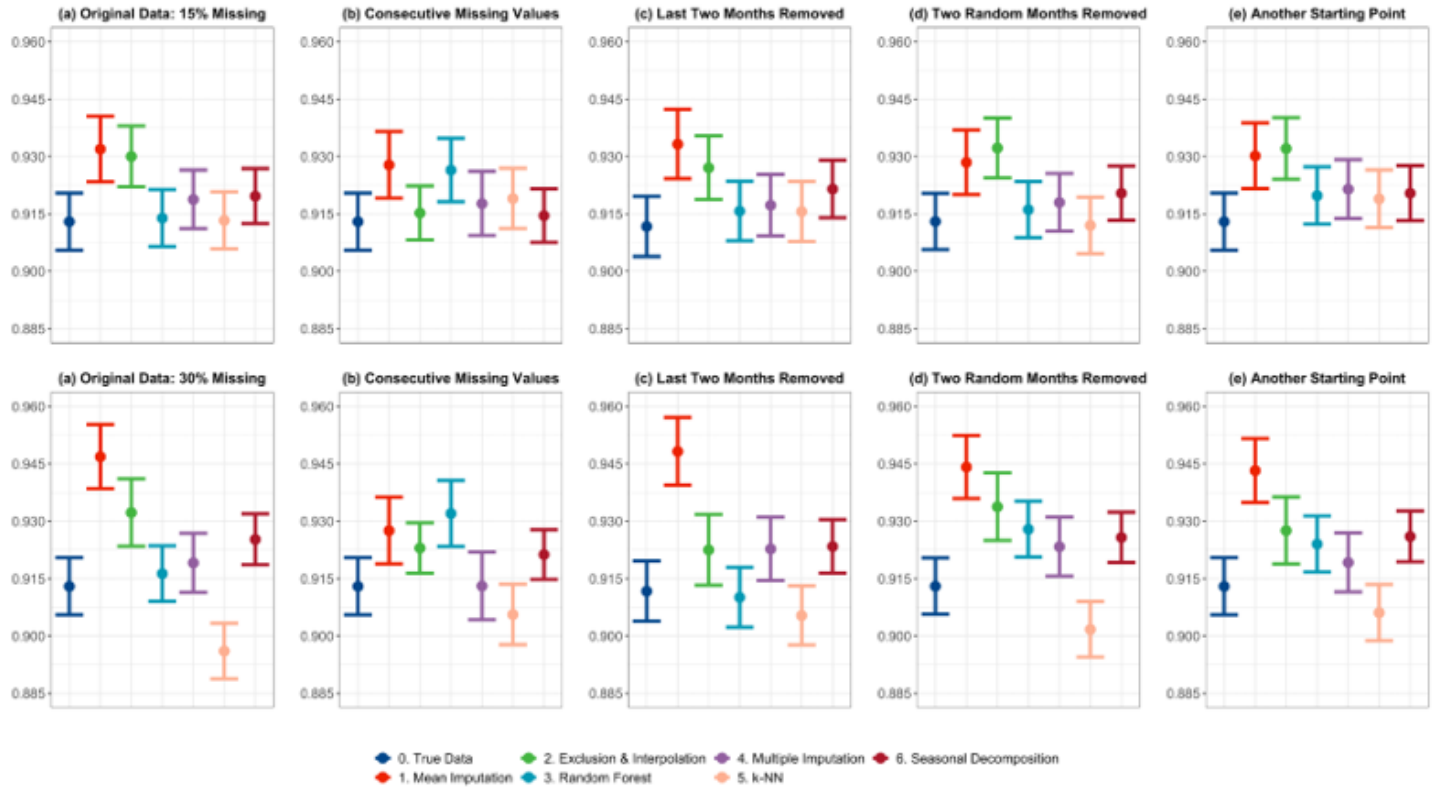

Figure 8

Estimated level change IRRs and 95% C.I.s with missing values inserted under different scenarios

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf
- SupplementaryMaterialsLegends.docx