Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

# Implementing Multiple Imputation for Missing Data in Longitudinal Studies When Models are Not Feasible: An Example Using the Random Hot Deck Approach

Chinchin Wang[1,2], Tyrel Stokes[3], Russell J Steele[3], Niels Wedderkopp[4], Ian Shrier[1]

[1]Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, Quebec, H3T 1E2, Canada; [2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, H3A 1A2, Canada; [3]Department of Mathematics and Statistics, McGill University, Montreal, Quebec, H3A 0B9, Canada; [4]Orthopedic Department University Hospital of South West Denmark, Department of Regional Health Research, University of Southern Denmark, Odense, Denmark

Correspondence: Ian Shrier, Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, 3755 Côte Ste-Catherine Road, Montreal, Quebec, H3T 1E2, Canada, Tel +1-514-229-0114, Email ian.shrier@mcgill.ca



**Purpose:** Researchers often use model-based multiple imputation to handle missing at random data to minimize bias. However, constraints within the data may sometimes result in implausible values, making model-based imputation infeasible. In these contexts, we illustrate how random hot deck imputation can allow for plausible multiple imputation in longitudinal studies.

**Patients and Methods:** Our motivating example is the Childhood Health, Activity, and Motor Performance School Study Denmark (CHAMPS-DK), a prospective cohort study that measured weekly sports participation for 1700 Danish schoolchildren. Using observed data on 4 variables (pain, activity frequency, sport, sport counts), we created a gold-standard data set without missing data. We then created a synthetic data set by setting some variable values to missing based on a prediction model that mimicked real-data missingness patterns. To create 5 imputed data sets, we matched each record with missing data to several fully observed records, generated probabilities from matched records, and sampled from these records based on the probability of each occurring. We assessed variability and agreement (kappa) between the imputed data sets and the gold-standard data set. We compare results to common model-based imputation methods.

**Results:** Variability across data sets appeared reasonable. The range of kappa for the random hot deck approach was moderate for activity frequency (0.65 to 0.71) and sport (0.59 to 0.85), and poor for common model-based approaches (range 0.00 to 0.11). The range of kappas for sport count was strong (0.87 to 0.97) for random hot deck imputation and weak to moderate (0.55 to 0.71) for common model-based imputation. Agreement was higher when more information was present, and when prevalence was higher for our binary variable sport.

**Conclusion:** Random hot deck imputation should be considered as an alternative method when model-based approaches are infeasible, specifically where there are constraints within and between covariates.

**Keywords:** multiple imputation, missing data, missing at random, hot deck imputation, random hot deck imputation, longitudinal studies

## Introduction

Missing data are present in most epidemiologic studies.[1] How they are handled is critical, as results may be biased and precision overestimated if inappropriate methods are used.[2] This article illustrates the use of random hot deck imputation to multiply impute unbiased values with appropriate confidence interval coverage when model-based methods are infeasible due to constraints between variables.

As a motivating example, the Childhood Health, Activity, and Motor Performance School Study Denmark (CHAMPS-DK) used sport participation data that contained multiple constraints between variables, which make standard model-based multiple imputation infeasible.[3] For instance, activity frequency must be greater than or equal to the number

Received: 30 March 2022
Accepted: 9 November 2022
Published: 15 November 2022

of sports played, and individual sport frequencies must equal the total activity frequency. Similar constraints in other studies might include side effects constrained by drug types or symptoms constrained by disease.

Appropriate methods for imputation depend on the intended analysis and the nature of missingness. Missing data fall into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are MCAR when missingness is independent of both observed and unobserved variables. Data are MAR when the missingness is only associated with observed variables. Data are MNAR when missingness is associated with unobserved variables.[4]

When estimating a causal effect with missing data, a simple way to handle missing data is to only analyze entries with complete data, ie complete case analysis. Complete case analysis is generally unbiased when data are MCAR,[5] or when the exposure or outcome is unrelated to missingness. MCAR is often an unreasonable assumption due to factors associated with both missingness and the outcome.[6] For instance, data collection might be more difficult in participants who are sicker and more likely to die.

Researchers typically work under the weaker assumption that data are MAR. Most established methods to handle MAR data without bias generally use model-based imputation, where models are used to predict missing values based on observed data.[7,8] MNAR data require advanced methods with complex assumptions and are beyond the scope of this article.[4]

In addition to bias, researchers should account for the greater uncertainty with imputed values compared to observed values.[9] Single imputation methods impute one replacement value for each missing value, so they cannot account for uncertainty in the imputed value. This results in confidence intervals that are too narrow.[8] Multiple imputation methods account for increased uncertainty by imputing with random variation to create several complete datasets. Each dataset is analyzed, and results are averaged,[8] providing more appropriate confidence intervals than single imputation.[9]

Most multiple imputation approaches are model-based. Briefly, a probability model is assumed that allows one to predict missing values by borrowing information on observed relationships between covariates in the data.[7,8] Regression parameters and/or imputed values are sampled from an underlying distribution[10] to generate different imputed values in multiple different data sets. This is generally straightforward for continuous data and implemented in most statistical software.[11]

Although model-based multiple imputation is generally preferable, there are contexts where implementing a model-based approach is infeasible due to strict dependencies amongst covariates, which impose constraints. In our motivating example for activity in the CHAMPS-DK study, if the number of activities is zero, the imputed value for total time spent must be exactly zero. In these and more complex constraint cases described later in this paper, common multiple imputation strategies implemented in standard software packages will often impute implausible values. Although more advanced techniques may sometimes solve the challenge, they may not be available to applied epidemiologists with less resources. Other options include using complete case analysis which may inappropriately assume MCAR and introduce bias,[12] or single imputation approaches which do not account for uncertainty of imputed values and may also introduce bias.[9,12] An alternative is to impute data using logic-based approaches that borrow information from observed data that already respects the imposed constraints.[13]

In this article, we propose a framework for applying the random hot deck imputation method[13] to missing clustered longitudinal data with complex constraints. Using the CHAMPS-DK data, we illustrate how to implement the framework for multiple imputation, with recommended methods for calculating appropriate confidence intervals.[13,14] To evaluate how well the method works, we used plasmode simulation methods.[15,16] In brief, we created a gold-standard data set based on observed values from the real-world data. We then created a synthetic data set with missing data based on the gold standard data (see Appendix). Next, we used the random hot deck method to multiply impute five data sets, and compared these to the gold-standard data set with respect to variability and level of agreement. Finally, we compared the agreement from our random hot deck approach to those obtained using common model-based imputation methods accessible to most epidemiologists.

# Methods

## General Approach for Random Hot Deck Imputation

Random hot deck imputation is a logic-based approach that uses rules to match missing records to sets of observed records to form a "donor pool". These rules can be developed to incorporate constraints between variables. We then randomly select one record from this donor pool for imputation.[13] This is in contrast to model-based imputation, where formal conditional models are used to predict missing values.[13]

By using a random mechanism for imputation, random hot deck imputation can incorporate the additional uncertainty surrounding missing values.[13] A similar method that combines model- and logic-based approaches is fully conditional specification, where missing variables are imputed one at a time using regression models with constraints specific for that variable.[17] However, it is difficult to apply to clustered longitudinal studies due to the larger number of constraints that must be incorporated into each regression model.[17] Incorporating constraints into random hot deck imputation is simpler and may be better suited for clustered longitudinal data.[13]

The five steps to apply random hot deck imputation and obtain appropriate confidence interval coverage are explained in Table 1. In brief, we choose covariates as we would in model-based imputation, identify a preliminary donor pool of records that match on the observed covariates within a missing record, increase uncertainty using Approximate Bayesian Bootstrap methods, derive sampling probabilities, and then create imputed data sets using the sampling probabilities.

## Motivating Example – CHAMPS-DK Study

We used data from the CHAMPS-DK study,[3] which was approved by the Ethics Committee for the region of Southern Denmark (ID S20080047). We focus on weekly data collected via SMS on children's pain and sport participation. If no response was received for a question, the next question was not asked. As SMS messages were sent in a free-text field, responses did not always contain clear answers for the variable of interest. Where possible, entries were corrected by deduction, or else coded as missing. These data included important constraints that prevented using standard imputation software for model-based imputation.

**Table 1** Five Steps Required for Random Hot Deck Imputation with Appropriate Confidence Interval Coverage

| Step | Explanation |
|---|---|
| 1. Identify covariates for matching that are related to the missing variable and the nature of their relationship, including possible constraints. | This step is equivalent to the identification of relevant covariates in model-based imputation. The covariates may be other variables, or as is typical in longitudinal data, the variable of interest measured at different time points. One should consider whether there are important time trends (eg seasonality) in the variable of interest or covariates. If so, one may decide to restrict the donor pool to records during specific time points in Step 2. |
| 2. Identify a preliminary (potential) donor pool by matching the record with missing data to other records based on observed covariate relationships, and the relevant time points given the context of the study. | It is important for this process to yield multiple matching records. If the criteria are such that only match is retrieved, then each imputed data set will have the same value, not allowing for multiple imputation. |
| 3. Create a final (actual) donor pool by resampling with replacement from the preliminary donor pool. | There is a higher risk of over-matching when using random hot deck imputation compared to model-based methods, and not accounting for over-matching can underestimate uncertainty.[15,16] This is resolved using the Approximate Bayesian Bootstrap.[15,16] Consider that we identified 3 records in our preliminary donor pool. If the preliminary donor pool had three records with values 1, 2, and 3, our final donor pool might be {1, 1, 2} for the first data set, {1, 3, 3} for the second data set, {1, 2, 3} for the third data set and so on. |
| 4. Derive sampling probabilities for replacement values based on the observed data for records within the final donor pool. | The sampling probabilities will depend on whether one is imputing based on covariates that are within-subject, between-subject or both. Concreate examples of both are provided in the Results section. |
| 5. Impute a replacement value according to the sampling probabilities for the final donor pool derived in Step 4. | Because we randomly sample with replacement, each of the imputed data sets will have different imputed values for the missing data. |

### Pain

Parents received an automated message asking whether their child experienced pain in their upper extremity, lower extremity, and spine in the past week, and whether pain was associated with a new injury (new pain) or continuing from a previous injury (old pain). These responses were converted into a composite variable (no pain, new pain in at least one body location, old pain in at least one body location and no new pain). Although there were missing data on pain, these could be imputed using model-based approaches. We use pain as a covariate when imputing other variables.

### Activity Frequency

After responding about pain, parents were asked to indicate the number of organized activity sessions (1–7, with 8 representing 8 or more) the child partook in outside of school that week.

### Types of Sports

After responding about frequency, parents were asked to indicate which sports were played in these sessions, with 1–9 representing different sports and 10 representing "Other". We refer to the number of times a child played each sport in a week as the "sport count".

### Sport Counts

How parents indicated which sports were played sometimes resulted in missing data on sport counts. Consider a child with an activity frequency of 4 because they played football (code 1) three times and handball (code 2) once. Parents might answer 1112, providing one sport for each activity session, with no missing sport count data. However, other parents might answer 12 without specifying how many sessions of each sport were played, resulting in missing sport counts.

There are multiple constraints amongst these variables that present challenges for model-based approaches. The activity frequency must be greater than or equal to the number of sports; each sport played must have an integer-valued sport count greater than 0; each sport not played must have a sport count of 0; and the sum of sport counts must total the frequency.

## Simulations to Evaluate Imputations

As a proof-of-concept objective, our aim was to compare the agreement between the gold standard data (oracle) and imputed data sets based on random hot deck methods, with common model-based multiple imputation methods (with and without constraints) that might be used by an applied epidemiologist who does not have special expertise in multiple imputation. To assess our random hot deck imputation approach, we generated our gold standard data using observed data from CHAMPS-DK and applying plasmode simulation methods.[15,16] Detailed methods and results are provided in the Appendix, including mean frequencies of sports played (Table S1). In brief, we created a gold standard dataset using complete records from the CHAMPS-DK dataset over a 26-week period. Using the gold-standard dataset, we created a synthetic dataset by randomly setting some records to be missing for frequency, sport, and sport count using a prediction model that mimicked the monotonic missingness patterns in the real data (Table S2, Figure S1). When missingness is not monotonic, random hot deck imputation is more challenging. These more complicated methods are described elsewhere.[13]

Our random hot deck imputation data sets were created using Approximate Bayesian Bootstrap methods to obtain appropriate uncertainty in the imputed five datasets. We assessed variability between imputed datasets. For frequency, we visually compared the distribution of activity frequency across the gold standard dataset and imputed datasets. For sport, we compared the number of weeks where each sport was played at least once across datasets. For sport count, we compared the mean sport counts in each week for each sport across datasets.

We implemented a common model-based imputation approach that is similar to what would be used by an applied epidemiologist without access to more advanced methods. We used the MICE package[18] with default settings in R for Statistical Computing.[19] We did not use a hierarchical model that would account for repeated measures on individuals because such models require specific expertise in multiple imputation. For the unconstrained model, we included identifier variables for each individual, as well as calendar week, sex, grade, school, pain, activity frequency, and sport. For the constrained method, we applied passive imputation methods[20] to ensure that the sum of all sport counts equaled the activity frequency. In some cases, the imputed activity frequency exceeded the maximum value of the gold standard data (8 = "8 or more activities"). In a sensitivity analysis, we truncated activity frequencies to a maximum value of 8 to ensure fair comparisons.

To assess agreement between the imputed datasets from both random hot deck and common model-based approaches against the gold standard dataset, our performance measures were Cohen's weighted kappa with quadratic weights for frequency and sport count and Cohen's unweighted kappa for sport. Primary calculations were performed only on imputed entries and the equivalent entries in the gold standard dataset, excluding non-imputed entries which would overestimate agreement. To gain further insight, we conducted the following analyses both overall and by missingness pattern for each of the binary sport variables (coded yes/no). We calculated the percentage of entries with correct imputations as described in the Appendix (Assessment of agreement section). In addition, we calculated the sensitivity (probability of correctly imputing a sport given that it was actually done) and specificity of imputation for each sport (probability of not imputing a sport given that it was actually not done).

## Results

We illustrate the proposed framework for applying random hot deck imputation for each of the variables activity frequency, type of sports, and sport count in the CHAMPS-DK data. Table 2 provides an overview of our approach.

**Table 2** Summary of Random Hot Deck Approach for Frequency, Sport, and Sport Count Variables. The Steps from Preliminary Donor Pool to Final Donor Pool and Applying the Sampling Method to the Data to Obtain the Imputed Value are the Same for Any Context or Study and are Omitted for Clarity

| Variable | Identify Covariates | Constraints | Preliminary Donor Pool | Derive Sampling Probabilities | Alternative Options |
|---|---|---|---|---|---|
| Frequency | • Pain<br>• Individual characteristics (eg gender, general level of activity)<br>• Age<br>• Seasonality<br>• Gender<br>• External factors (eg school events, weather) | • Not applicable | • Match on pain within individuals in nearby weeks (accounts for within individual characteristics, age, and seasonality) | • Sample difference between individual frequency and gender-specific median class frequency from donor pool<br>• Add difference to the median class frequency for the missing week (accounts for age, gender, and external factors between individuals) | • Sample frequency directly from donor pool (does not account for external factors) |
| Sport | • Individual characteristics (eg gender, sport preference)<br>• Age<br>• Seasonality<br>• Frequency | • Number of sports cannot be greater than frequency | • Match on nearest frequency within individuals in nearby weeks (accounts for individual characteristics, age, and seasonality) | • Sample sport from donor pool<br>• If number of sampled sports is greater than the frequency, sample with replacement an equal number of sports as the frequency based on their relative proportion in nearby weeks (accounts for seasonality and frequency) | • Additionally match on pain (reduces number of matching records within chosen time frame) |
| Sport Count | • Individual characteristics (eg gender, sport preference)<br>• Age<br>• Seasonality<br>• Sport<br>• Frequency | • Number of sport counts must equal frequency<br>• Sport counts must be >0 for all sports played<br>• Sport counts must be 0 for all sports not played | • All nearby weeks within individuals where at least one of the sports were played (accounts for individual characteristics, age, seasonality, and sport) | • Calculate relative proportion of each sport in nearby weeks<br>• Sample with replacement an equal number of sports as the frequency based on their relative proportions (accounts for frequency) | • Assume some sports are more likely to be played in the same week (more complex logic) |

# Imputing Activity Frequency

## Identify Covariates for Matching

One's activity frequency in a particular week is likely influenced by the presence or absence of pain. Additionally, activity frequency tends to change with season and age, and is likely more similar in nearby weeks than weeks further away. Further, children in the same class at school are exposed to similar factors (eg weather, school events) that may lead to individuals of the same class and gender being more active in certain weeks. To account for these external factors, we considered gender-specific median class frequencies in the missing and nearby weeks as covariates.

## Identify a Preliminary Donor Pool

We generated our preliminary (potential) donor pool by matching within individuals on pain in nearby weeks. We used our composite pain variable with 3 levels (no pain, new pain, old pain), assuming pain in different body locations has the same effects on activity frequency.

In our context, we believed a timeframe of 7 weeks before and after the missing week was appropriate because it encompasses weeks of the same season (3-month period), and we assume participants are more likely to be active in some seasons compared to others (Figure 1A and B). One alternative might be to limit the timeframe to specific dates such as the start and end of a season for a particular organized sport. These are value choices, and investigators might consider sensitivity analyses to assess the influence of these decisions.

When no matches were available in the 7 weeks before and after (3-month period), we extended the pool to include entries 12 weeks before and after (6-month period), 25 weeks before and after (1-year period), then the entire study. Others might choose smaller time windows when covariate relationships are sensitive to changes over time. Occasionally, no matches existed because the individual did not have any other entries with a particular pain value (new or old). However, they may have had entries with the other pain value. In these cases, we matched on any pain (new or old) in the 7 weeks before and after, and so forth. In cases where the individual had no entries with pain except the missing entry, we sampled from all entries in the 7 weeks before and after, even though these weeks had no pain.

## Create Final Donor Pools Using Approximate Bayesian Bootstrap

We sampled the preliminary donor pool from Step 2 with replacement (number of samples equal to the number of matched records) to create additional uncertainty in the final (actual) donor pool. To create five imputed data sets, this procedure is conducted 5 times. For simplicity of presentation, we chose a final donor pool for our figures that matched the preliminary donor pool. Note that the Approximate Bayesian Bootstrap method would sometimes produce final donor pools that are identical to preliminary donor pools as we have chosen. However, the method will often produce different final donor pools where one or several rows from the preliminary donor pool appear multiple times, ie the final and preliminary donor pools are usually different.

## Derive Sampling Probabilities

Each week in the final donor pool has an equal probability of being sampled. To create each imputed data set, we randomly sampled one week from each of the final donor pools created (Figure 1C). We first imputed the difference between the individual's frequency and their gender-specific median class frequency from the sampled week. This was used as a measure of how much activity they did relative to their peers in the missing week to account for external factors (eg school events, weather) causing similar individuals (of the same class and gender) to change their activity. The imputed frequency for the missing week was the sum of the gender-specific median class frequency in the missing week and this difference (Figure 1D).

Our procedure is similar to random generation of values from a fixed effects model. We estimate the fixed effect for activity frequency from the observed data (the median frequency for individuals of the same class and gender, whom we assume come from the same distribution of relevant background characteristics). We sample a residual (the difference between the individual's frequency and the median frequency for their class and gender) from the final donor pool of potential matches. The imputed value is then the sum of the fixed effect and the residual.

**A**

| Pain | Frequency | Median Class Frequency | Freq minus Median Class Freq |
|---|---|---|---|
| No pain | 2 | 1 | 1 |
| New pain | 3 | 2 | 1 |
| Old pain | 0 | 2 | -2 |
| Old pain | 1 | 1 | 0 |
| Old pain | 0 | 1 | -1 |
| No pain | 1 | 0 | 1 |
| No pain | 1 | 1 | 0 |
| No pain | | 2 | |
| No pain | 2 | 1 | 1 |
| Old pain | 1 | 1 | 0 |
| No pain | 0 | 0 | 0 |
| No pain | 2 | 1 | 1 |
| New pain | 2 | 1 | 1 |
| No pain | 3 | 2 | 1 |
| No pain | 3 | 2 | 1 |

**B**

| Pain | Frequency | Median Class Frequency | Freq minus Median Class Freq |
|---|---|---|---|
| No pain | 2 | 1 | 1 |
| New pain | 3 | 2 | 1 |
| Old pain | 0 | 2 | -2 |
| Old pain | 1 | 1 | 0 |
| Old pain | 0 | 1 | -1 |
| No pain | 1 | 0 | 1 |
| No pain | 1 | 1 | 0 |
| No pain | | 2 | |
| No pain | 2 | 1 | 1 |
| Old pain | 1 | 1 | 0 |
| No pain | 0 | 0 | 0 |
| No pain | 2 | 1 | 1 |
| New pain | 2 | 1 | 1 |
| No pain | 3 | 2 | 1 |
| No pain | 3 | 2 | 1 |

**C**

| Pain | Frequency | Median Class Frequency | Freq minus Median Class Freq |
|---|---|---|---|
| No pain | 2 | 1 | 1 |
| New pain | 3 | 2 | 1 |
| Old pain | 0 | 2 | -2 |
| Old pain | 1 | 1 | 0 |
| Old pain | 0 | 1 | -1 |
| No pain | 1 | 0 | 1 |
| No pain | 1 | 1 | 0 |
| No pain | | 2 | *1* |
| No pain | 2 | 1 | 1 |
| Old pain | 1 | 1 | 0 |
| No pain | 0 | 0 | 0 |
| No pain | 2 | 1 | 1 |
| New pain | 2 | 1 | 1 |
| No pain | 3 | 2 | 1 |
| No pain | 3 | 2 | 1 |

**D**

| Pain | Frequency | Median Class Frequency | Freq minus Median Class Freq |
|---|---|---|---|
| No pain | 2 | 1 | 1 |
| New pain | 3 | 2 | 1 |
| Old pain | 0 | 2 | -2 |
| Old pain | 1 | 1 | 0 |
| Old pain | 0 | 1 | -1 |
| No pain | 1 | 0 | 1 |
| No pain | 1 | 1 | 0 |
| No pain | *2 + 1 = **3*** | 2 | *1* |
| No pain | 2 | 1 | 1 |
| Old pain | 1 | 1 | 0 |
| No pain | 0 | 0 | 0 |
| No pain | 2 | 1 | 1 |
| New pain | 2 | 1 | 1 |
| No pain | 3 | 2 | 1 |
| No pain | 3 | 2 | 1 |

**Figure 1** Imputation of activity frequency. (**A**) There is one week where frequency is missing (black row). Pain is coded as no pain in any location (No pain), new pain in at least one location (New pain), and old pain in at least one location but no new pain (Old pain). The individual had no pain in the week with missing data. The median frequency for the individual's class and gender is calculated for the missing and surrounding weeks (Median Class Frequency). For the weeks with observed data, we also calculate the difference between the individual's frequency and the median frequency (Freq minus Median Class Freq) as a measure of how much activity the individual does relative to their class and gender. (**B**) We match on nearby weeks with the same level of pain (gray rows). The preliminary donor pool is comprised of eight weeks where the individual also experienced no pain. (**C**) For simplicity of presentation, we chose a final donor pool that happened to exactly match the preliminary donor pool in (**B**). One of the weeks in the final donor pool is randomly selected (outlined in black). The difference between the individual's frequency and the median class frequency for the sampled week is 1. This difference is imputed for the missing week. (**D**) The imputed frequency for the missing week is the sum of the median class frequency for the missing week and the imputed difference between the individual and median class frequency. In this example, the imputed difference of 1 is added to the median class frequency of 2 to obtain an imputed frequency of 3.

## Impute Replacement Value

We applied the derived sampling probability to each of the final donor pools and imputed the selected value for each imputed data set.

### Simulation Results for Frequency

Variability. Figure S2 displays the number of observations for each activity frequency in the gold standard dataset and five imputed datasets for random-hot deck and model-based imputations, restricted to missing entries in the

synthetic dataset (2% of dataset). Due to the randomness of the imputation method, each imputed dataset had a slightly different activity frequency distribution. In the data sets, imputation of individual sport counts sometimes resulted in a sum of sport counts greater than 8. In the constrained data set, the activity frequency had to equal the sum of all activities, resulting in values for activity frequency that were greater than 8. However, activity frequency had a maximum value of 8 in the gold standard (8 = 8 or more activities). In a sensitivity analysis, we truncated activity frequency to a maximum value of 8 in order to make a fair comparison.

**Agreement.** When comparing all entries in the gold standard to the imputed datasets, the weighted Cohen's kappa for frequency ranged between 0.98 and 1.00 for random hot deck imputation. This is an overestimate due to the large amount of non-imputed data for this variable (98% of entries). When restricted to missing entries in the synthetic dataset, the average weighted Cohen's kappa for frequency was 0.67 (range: 0.65 to 0.71) for random hot deck imputation (moderate agreement[21]), 0.02 (range: −0.05 to 0.10) for the common unconstrained model-based imputation and 0.00 (range: −0.06 to 0.09) for common constrained model-based imputation (no agreement).

## Imputing Sport

### Identify Covariates for Matching

We assume that while individuals are likely to play similar sports in nearby weeks, these sports might change over time and season. If an individual never played a particular sport, we assume that sport was not played in a missing week. We also assume that individuals played similar sports with similar frequencies in nearby weeks.

### Identify a Preliminary Donor Pool

We generated our preliminary donor pool by matching within individuals on closest frequency to the missing week within nearby weeks (7 weeks before and after) (Figure 2A and B). We did not extend the time window to match on exact frequency because as opposed to activity frequency, we assumed that sports participation is very likely to differ by season. This is an example where constraints in our data make model-based approaches difficult. Again, an alternative might be to set the timeframe so that the start and end coincide with the start and end of an organized sport season.

If there were no matches on closest frequency (ie if an individual did not have any observed frequency in nearby weeks or all frequencies were 0), we extended the time window to 12 weeks before and after (6-month period), 25 weeks (1-year period), then the entire study.

### Create Final Donor Pools Using Approximate Bayesian Bootstrap

We sampled the preliminary donor pool from Step 2 with replacement to create additional uncertainty in the final donor pools. For simplicity, the final donor pools in our figures exactly match the preliminary donor pool.

### Derive Sampling Probabilities

Each week in the final donor pool had an equal probability of being sampled. If all records in the final donor pool had the same frequency as the missing week, we randomly sampled one of these weeks (Figure 2C) and imputed its sports (Figure 2D).

Sometimes the final donor pool contained records with frequencies lower or greater than the missing week. If the sampled week had a lower frequency than the missing week, we imputed its sports. These records would still be missing sport counts (number of times each sport was played). We describe how we imputed sport counts in Section 2.5.

If the sampled week had a frequency greater than the missing week, two possibilities existed. If the number of sports was equal or less than the frequency of the missing week, we imputed the sampled week's sports. However, if the number of sports was greater than the frequency of the missing week, imputing all the sports would break our constraints. In Figure 3A, the missing week has a frequency of 2, but nearby weeks have frequencies of 3 or more. The sampled week (Figure 3B) has three sports. To determine which sports to impute, we calculated sampling

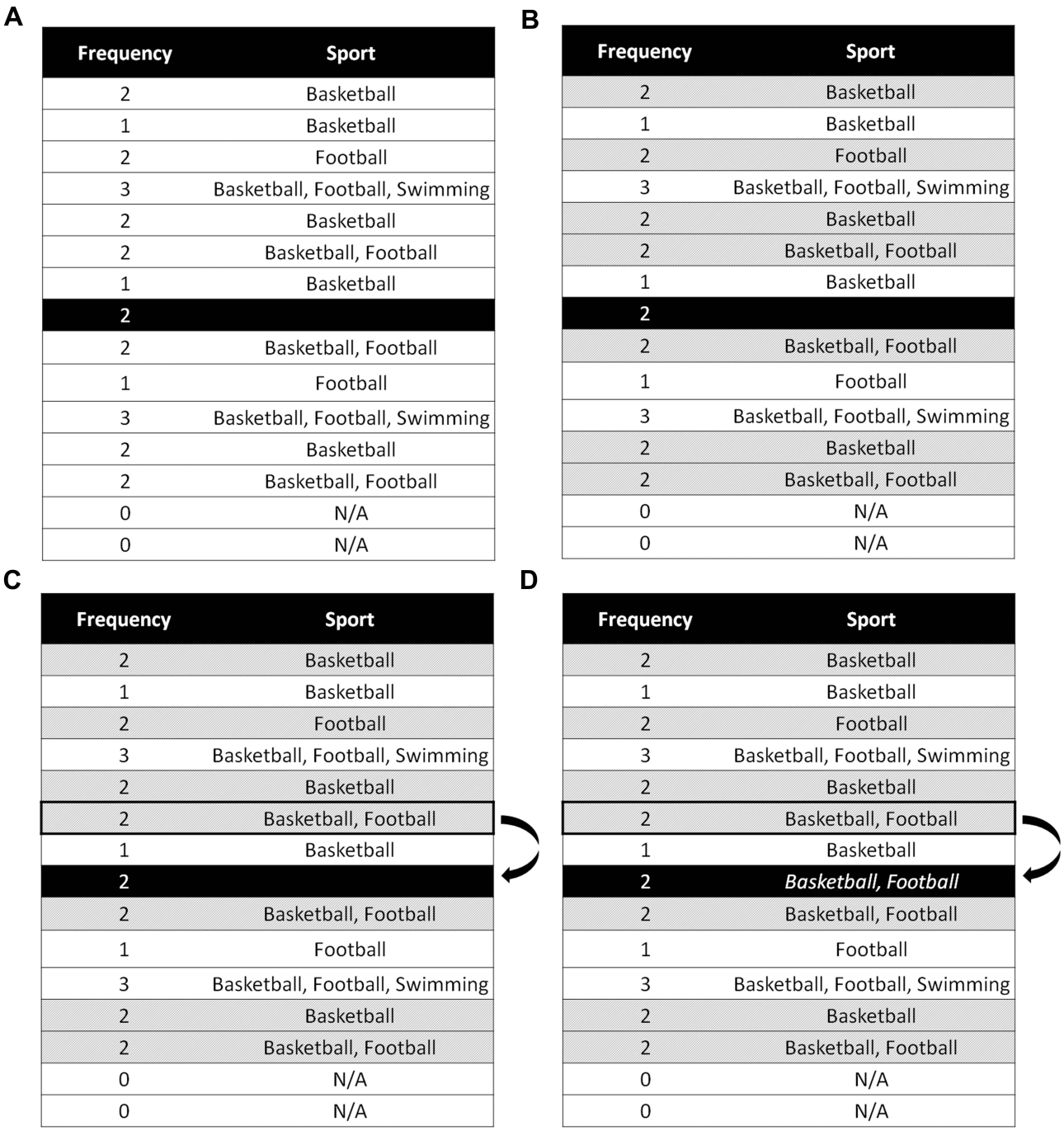


**Figure 2** Imputation of sport. (**A**) There is one week where the sports performed are missing (black row). The individual had a total activity frequency of 2 in this week. (**B**) We match on closest frequency in the nearby weeks. The preliminary donor pool is comprised of weeks where the individual also had frequencies of 2 (gray rows). (**C**) For simplicity of presentation, we chose a final donor pool that happened to exactly match the preliminary donor pool in (**B**). One of these weeks is randomly sampled with equal probability (outlined in black). (**D**) The sports from the sampled week are imputed for the missing week.

probabilities for each of the three sports according to their relative proportion in nearby weeks and sampled two sports with replacement (Figure 3C and D).

Our approach assumes that the sports one plays were only related to the individual's activity frequency that week and sports played in nearby weeks (ie seasonality). More complex matching constraints could include the presence of pain. More constraints will reduce the number of matching records within the chosen time frame.

## Simulation results

**Variability.** Figure S3 displays the number of entries where each sport was done in the gold standard and imputed datasets, restricted to missing entries in the synthetic dataset (3% of dataset). As with activity frequency, there was variation between the imputed datasets due to the randomness of the imputation method.

**Agreement.** When comparing all entries in the gold standard and imputed datasets, the unweighted Cohen's kappa for each of the binary sport variables using random hot-deck imputation ranged between 0.98 and 1.00. This is an overestimate due to the large amount of non-imputed data for these variables (97% of entries). When restricted to missing entries in the synthetic dataset, the average kappa across all records imputed for sport using random hot

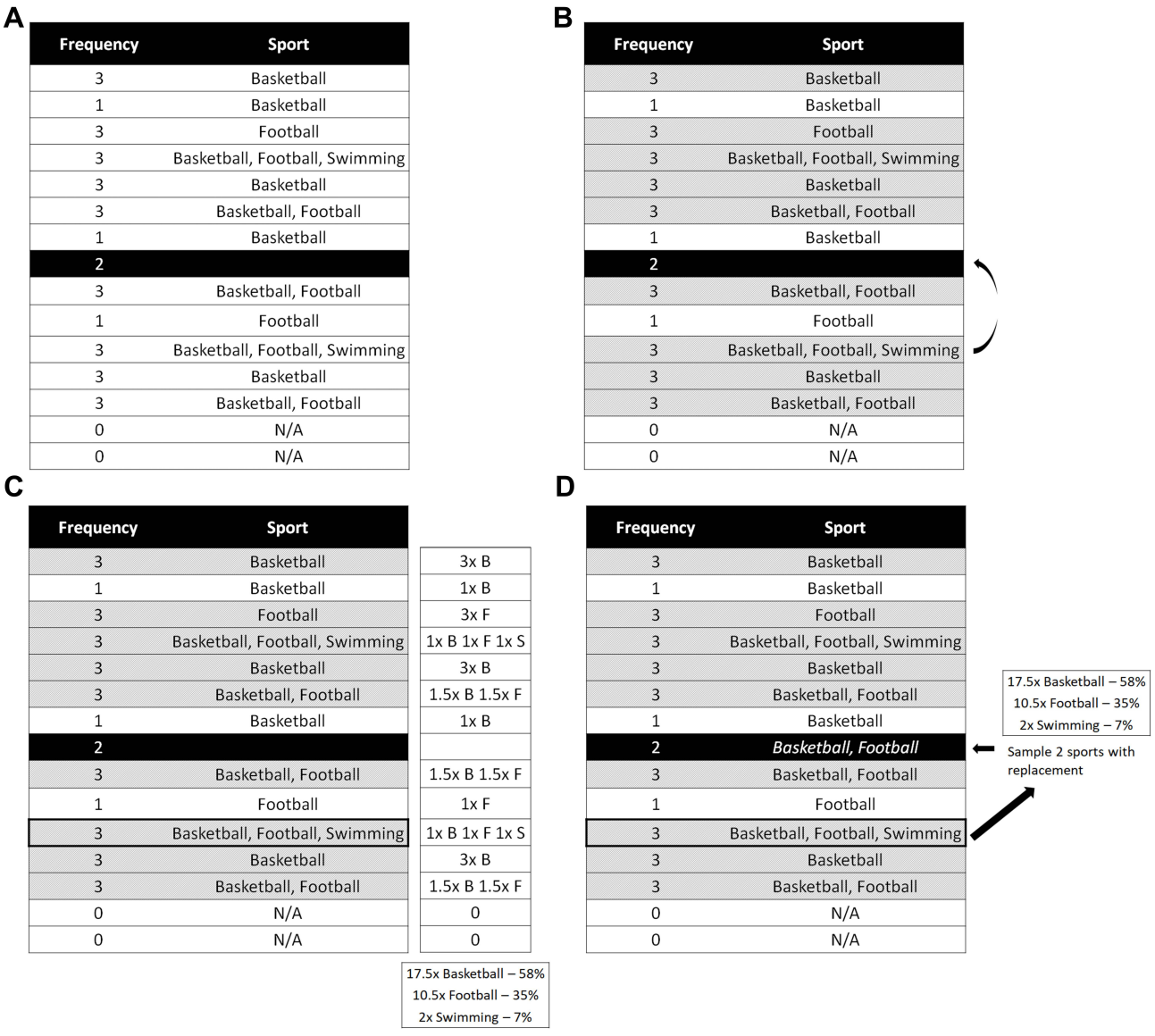


**Figure 3** Imputation of sport where the number of sports is greater than the frequency. (**A**) There is one week where the sports performed are missing (black row). The individual had a total activity frequency of 2 in this week. (**B**) The preliminary donor pool is comprised of nearby weeks with the closest frequency to the missing week. Since there are no weeks with a frequency of 2, we match on weeks with frequencies of 3 (gray rows). For simplicity of presentation, we chose a final donor pool that happened to exactly match the preliminary donor pool in (**B**). One of the matched weeks is randomly sampled. (**C**) The sampled week has 3 sports, while the missing week only has a frequency of 2. The number of times in nearby weeks that the individual participated in each sport is determined. For weeks where the frequency is greater than the number of sports, the frequency is divided equally. The relative amount that the individual participated in each sport in nearby weeks is used as the sampling probability. Since the individual did basketball 10.5 times, football 7.5 times, and swimming 1 time, the probabilities are 55% (10.5/19), 40% (7.5/19), and 5% (1/19) respectively. (**D**) Sports are randomly sampled using the sampling probabilities and imputed for the missing week. Basketball and football are randomly imputed.

deck imputation ranged from 0.59 to 0.85 depending on the sport (Figure 4) demonstrating moderate-to-strong agreement.[21] The corresponding average kappas using common model-based imputation without constraints ranged from −0.02 to 0.05 depending on the sport and 0.00 to 0.11 using the constrained common model-based imputation (data not shown). As expected, kappas were much higher for entries where frequency was observed but sport was missing, than when both frequency and sport were missing. This can be attributed to the additional information that frequency provides about sport. Kappas were generally lower for sports that were done less frequently, as is expected for sports with lower prevalence.[22]

Overall, missing sports were imputed correctly (imputing each sport that was played and imputing no sports that were not played) in 71% of entries using random hot deck imputation, 21% of entries using common model-based imputation, and 24% using common model-based imputation with our one constraint. When frequency data were available, missing sports were imputed correctly in 85% of entries using random hot deck imputation, 21% of entries using common model-based imputation without constraints and 38% of entries with our one constraint. When frequency data were missing, 65% of entries were imputed correctly using random hot deck imputation, 21% of entries using common model-based imputation without constraints and 18% with our one constraint. The lower values when frequency data were missing are due to the reduced information.

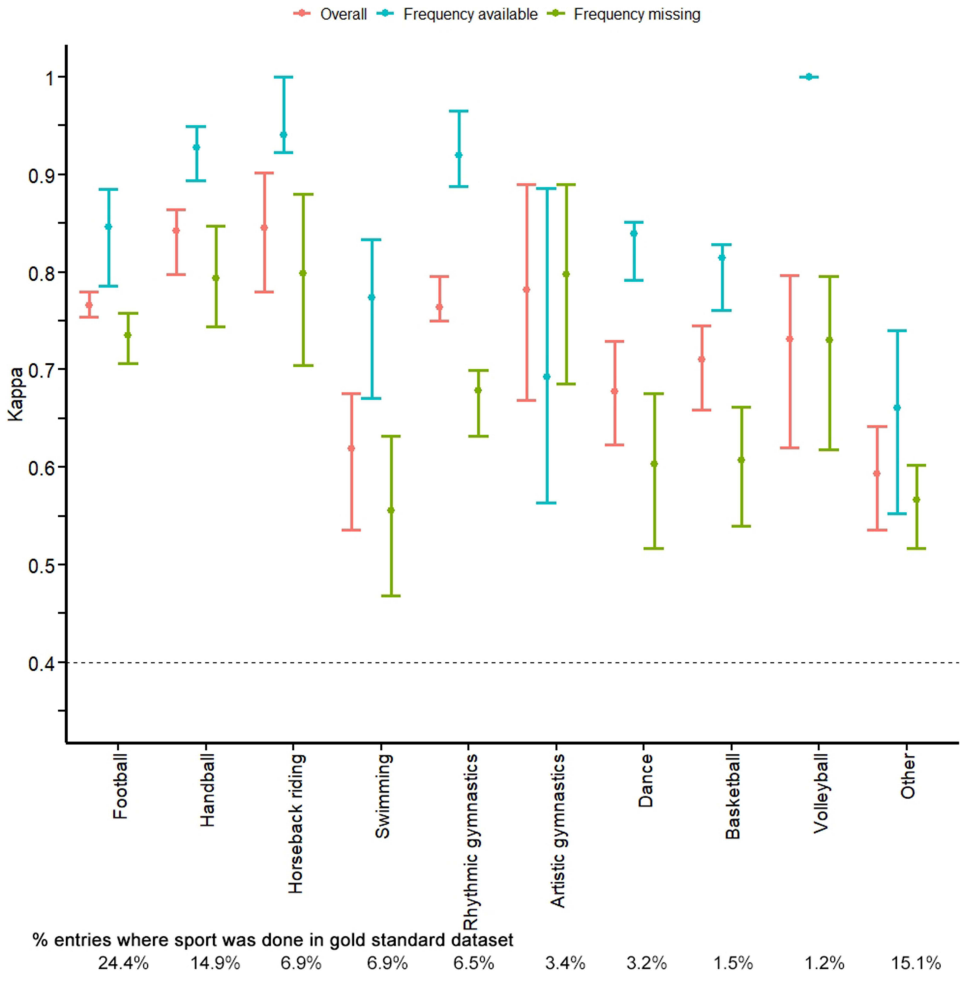


**Figure 4** Agreement between gold standard and imputed datasets for sport by missingness pattern. "Overall" is a weighted average of the two missingness patterns (frequency data available, frequency data missing). Agreement was assessed between the gold standard and imputed datasets using an unweighted kappa. Points represent the average kappa across 5 imputed datasets, while bars represent the range of kappas between the 5 imputed datasets. The horizontal dashed line represents the cut-off for minimal agreement (kappa = 0.40) (McHugh 2012, *Biochem. Medica.*). Calculations were restricted to missing entries in the synthetic dataset; kappas calculated using the entire dataset ranged from 0.98 to 1.00. The percentage of entries where each sport was done in the gold standard dataset is shown below the plot.

Overall sensitivity for the ten sports ranged from 0.60 to 0.92 (Figure S4) using random hot deck imputation. Sensitivity tended to be lower for sports done less frequently, and when frequency data were missing. Sensitivity for common model-based imputation ranged from 0.07 to 0.81 without constraints and 0.07 to 0.63 with our one constraint. Overall specificity ranged from 0.92 to 1.00 for the random hot deck imputation (Figure S5). Specificity for the common model-based imputation ranged from 0.79 to 0.99 without constraints and 0.74 to 0.99 with our one constraint. Specificity tended to be lower for sports done more frequently, and when frequency data were missing.

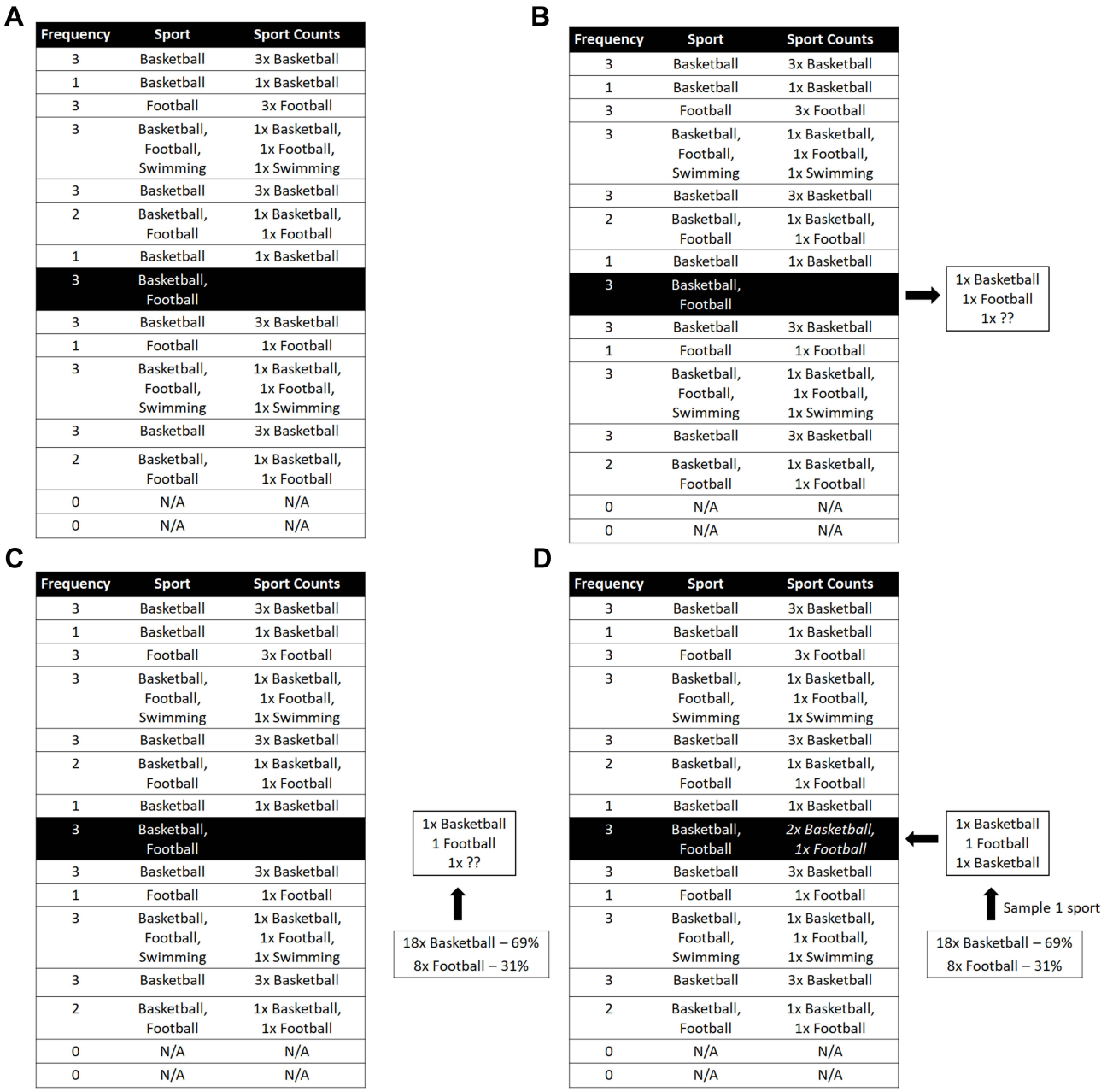


**Figure 5** Imputation of sport counts where a single week is missing. (**A**) There is one week where the total frequency is greater than the number of sports performed (black row). We would like to impute individual counts for each sport that was done. (**B**) The individual participated in at least one session of basketball and one session of football. As the total frequency for the missing week is 3, we still need to impute a single count that is either basketball or football. (**C**) For simplicity of presentation, we chose a final donor pool that happened to exactly match the preliminary donor pool in (**B**). The relative proportion of each sport done in the missing week (basketball and football in (**C**)) is calculated for the nearby weeks and used as the sampling probabilities. As basketball was done 9 times and football was done 5 times, the probabilities are 64% (9/14) and 36% (5/14) respectively. (**D**) Basketball is randomly sampled for the missing sport. Sport counts are imputed as two sessions of basketball and one session of football.

## Imputing Sport Counts

In this study, most parents indicated which sports were played but not how many times each sport was played (eg Figure 5A, frequency "3", sport "Basketball, Football"). We imputed sport counts in cases where the total frequency (observed or imputed) was greater than the number of sports. For each week with missing sport counts, each listed sport was played at least once. Therefore, the number of missing counts is the total frequency minus the number of sports (Figure 5B).

### Identify Covariates for Matching

Similar to activity frequency and sports participation, we believe the relative frequency with which individuals participate in different sports differs by age, gender and season. Individuals are likely to have similar relative frequencies in nearby weeks. Because individuals participate in different sports at different frequencies, we only borrowed information from within individuals.

### Identify a Preliminary Donor Pool

Our preliminary donor pool included all nearby weeks (7 weeks before and after) for the individual with missing sport counts that included sports from the missing week. Because sports that were not played must have a zero probability for the remaining counts, other weeks were not included in the pool.

### Create Final Donor Pools Using Approximate Bayesian Bootstrap

We sampled the preliminary donor pool from Step 2 with replacement to create additional uncertainty in the final donor pools. For simplicity, the final donor pools in our figures exactly match the preliminary donor pool.

### Derive Sampling Probabilities

Probabilities were derived by dividing the total counts for each sport in nearby weeks by the total frequency that the sports in the missing week were played in nearby weeks (Figure 5C). In our example, we know the child played basketball and football once each. In nearby weeks, they played basketball 18 times and football 8 times (69% basketball; 31% football). We used these probabilities to sample the remaining sport count so that the total sport count matched the frequency (Figure 5D).

Sometimes, sport counts were also missing for nearby weeks (Figure 6A). For these weeks, we divided the frequency by the number of sports to obtain average counts (Figure 6B). These counts were not imputed into the main dataset; rather, they were temporarily used to determine the probabilities for the week of interest. The counts from observed weeks and average counts for missing weeks were then summed as above and divided by the total frequency to obtain sampling probabilities (Figure 6C and D). Once data were imputed for the first missing week, we proceeded to the next missing week.

Our approach assumes that the probability of playing a particular sport is independent of the other sports conditional on the weekly sport count. Alternatively, we could use more complex logic that assumes some sports are more likely to be played together, and adjust our sampling scheme accordingly.

### Simulation Results

**Variability.** Figure S6 displays the mean sport count for each sport where that sport was done (sport count > 0) for the gold standard and imputed datasets, restricted to missing entries in the synthetic dataset (13% of the dataset). As for activity frequency and sport, the mean sport counts differed randomly between imputed datasets.

The variability is much less than we observed for Frequency and Sport because of the constraints in the data. In Table S2, we see that the total proportion missing sport counts is 12.8%. Of these, 77.3% (9.9/12.8) of these participants had information on frequency and sport. In addition, as we explained in Figure 6, a participant with Frequency = 3 who listed Basketball and Football as sports would only require imputing one of the sport counts and not all three sport counts. This constraint again reduces the variability in our imputed data sets compared to the variables Frequency and Sport. The differences between the gold standard data and the imputed data sets using the common model-based approach with and without constraints are much greater than that for the random hot deck approach.

Agreement. When comparing all entries in the gold standard and imputed datasets, the weighted Cohen's kappa for each of the sport count variables ranged between 0.98 and 1.00 for random hot deck imputation.

Figure 7 shows the average kappa ranged across sports from 0.87 to 0.97 for random hot deck imputation when restricted to missing entries in the synthetic dataset, demonstrating strong-to-almost perfect agreement.[21] The corresponding average kappa for common model-based imputation without constraints ranged from 0.55 to 0.71, demonstrating weak to moderate agreement, and ranged from −0.01 to 0.33 when we included our one constraint, demonstrating no

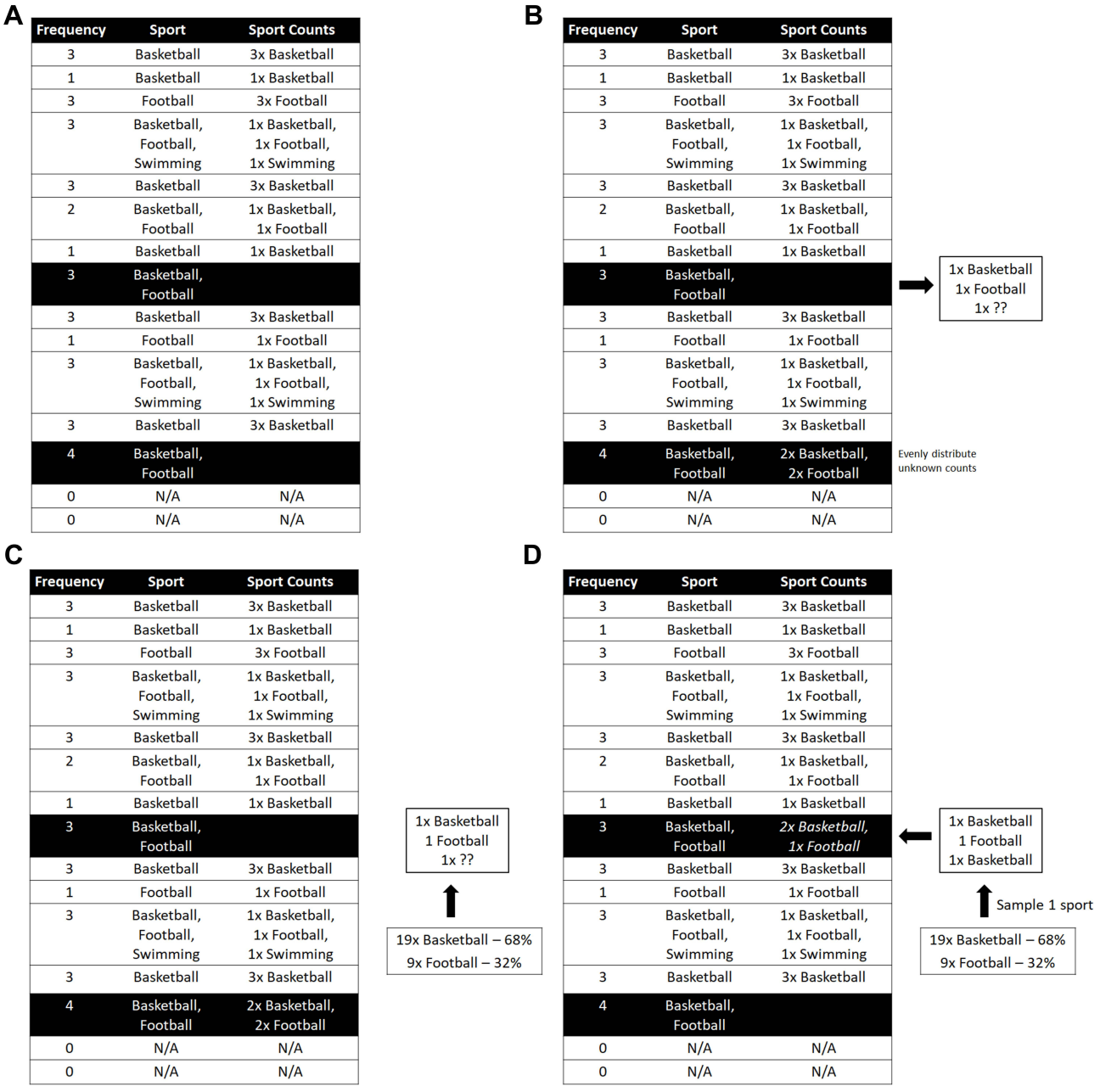


**Figure 6** Imputation of sport counts where multiple weeks are missing. (**A**) There are two weeks where the total frequency is greater than the number of sports (black rows). We would like to impute individual counts for each sport that was done for both weeks. We focus on imputing sport counts for the first missing week. (**B**) In the missing week, the individual had a frequency of 3 and participated in basketball and football. They must have participated in one session each of basketball and football. We must therefore impute a single count. Sport counts are calculated for the nearby weeks. When the frequency is greater than the number of sports (ie for the week with frequency 4), the counts are evenly distributed (assigning 2 counts to basketball and 2 counts to football). (**C**) For simplicity of presentation, we chose a final donor pool that happened to exactly match the preliminary donor pool in (**B**). The relative proportion of each sport in the final donor pool (ie matching the sports that were done in the missing week) is calculated for the nearby weeks and used as the sampling probabilities. Since basketball was done 10 times and football 7 times, the sampling probabilities are 59% (10/17) and 41% (7/17) respectively. (**D**) Basketball is randomly sampled. Sport counts are imputed as 2 sessions of basketball and 1 session of football.

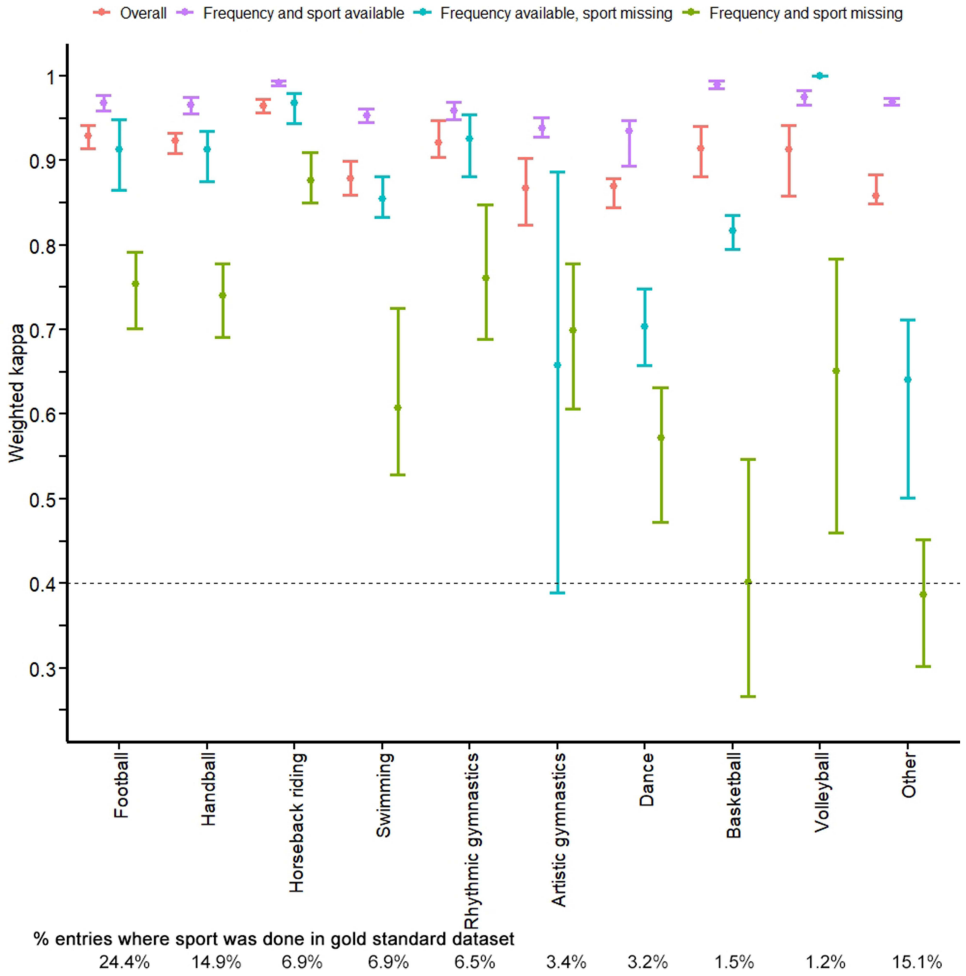


**Figure 7** Agreement between gold standard and imputed datasets for sport count by missingness pattern. "Overall" is a weighted average of the three missingness patterns (frequency and sport data available; frequency data available but sport data missing; frequency and sport data missing). Agreement was assessed between the gold standard and imputed datasets using a weighted kappa. Points represent the average kappa across 5 imputed datasets, while bars represent the range of kappas between the 5 imputed datasets. The horizontal dashed line represents the cut-off for minimal agreement (kappa = 0.40) (McHugh 2012, Biochem. Medica.). Calculations were restricted to missing entries in the synthetic dataset; kappas calculated using the entire dataset ranged from 0.98 to 1.00. The mean sport count for each sport in the gold standard dataset is shown below the plot.

to minimal agreement (data not shown). As with sport, agreement for sport count was higher when more information was present. For both random hot deck and model-based imputations, kappas were highest for entries where frequency and sport were observed, followed by when frequency was observed but sport was missing, followed by when frequency and sport were each missing. Kappas were generally lower for sports that were done less frequently, as is expected for sports with lower prevalence.[22] Kappas below 0.4 using random hot deck imputation only occurred for sports with low prevalence when each of frequency and sport were missing.

## Discussion

Our results illustrate that random hot deck imputation is an alternative to traditional model-based multiple imputation when developing a plausible model is infeasible. We obtained relevant variability across the imputed data sets, and the level of agreement was generally moderate, with lower levels of agreement as expected when prevalence was very low or we were not able to borrow much information from either between or within participants. The common model-based methods with and without our one constraint performed very poorly, presumably because they ignore within participant correlation and our random hot deck approach explicitly took this into account. Model-based imputation that accounts for repeated measures would likely perform better but require special expertise in imputation methods not available to many

epidemiologists. Even for statisticians with special expertise in imputation methods, our random hot deck approach could be used to assess the performance of the more advanced model-based methods.

While random hot deck imputation has commonly been used in surveys, several extensions to clustered longitudinal data have been described in the literature. Little et al[23] and Wang et al[24] applied this method to multiply impute gaps in recurrent event data, using menstrual patterns as an example. Unlike previous extensions which only borrowed information between individuals, our approach restricts the donor pool to entries between and within individuals where appropriate.

This approach uses the same data to impute missing values as a model-based approach would use. In a model-based approach, one tries to encode domain knowledge about the relationships between variables in a joint probability model or set of conditional probability models so that data are MCAR conditional on the covariates in the model. Here, we use domain knowledge to create hierarchical strata, also assuming MCAR conditional on the matching covariates. By resampling values within the strata under these conditions, random hot deck imputation produces consistent estimates conditional on all relevant information.[13] Choosing the number of imputations also follows the same principles as model-based approaches.[25]

The randomness in the approach also allows one to compute confidence intervals using standard rules that account for variations within and between datasets. However, after identifying a potential donor pool of records, one must increase the uncertainty in this donor pool across data sets using the Approximate Bayesian Bootstrap method to avoid over-matching and inappropriately narrow confidence intervals.[13,14] When sample sizes are small or percentage of responses is low, approaches other than the Approximate Bayesian Bootstrap method are recommended.[13]

While our motivating example was sport participation, our approach can be applied to any context with constraints between variables. For example, activity frequency is akin to the total number of medications in a pharmacoepidemiology study. Sports are akin to drug types, and sport counts to side effects. The number of drug types cannot exceed the total number of medications. Probabilities of various side effects differ between drugs, and some side effects may never occur for certain drugs.

## Conclusion

We recommend researchers use model-based multiple imputation where possible. In clustered longitudinal data, a model-based approach may be infeasible because of constraints between variables. In these contexts, random hot deck multiple imputation with Approximate Bayesian Bootstrap can provide less biased results and more appropriate levels of uncertainty than complete case analysis or single imputation methods. Although this approach requires many assumptions, any model-based approach would have to include the same assumptions or risk imputing implausible values.

## Data Sharing Statement

The datasets generated and/or analysed during the current study are not publicly available due to legal and ethical restrictions but are available from the CHAMPS Study Steering Committee (NWedderkopp@health.sdu.dk) on reasonable request. The code used for the worked examples is available on Open Science Framework (https://osf.io/nyd8x/).

## Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agreed to be accountable for all aspects of the work.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Disclosure

The authors declare that they have no competing interests.

## References


1. Perkins NJ, Cole SR, Harel O., et al. Principled approaches to missing data in epidemiologic studies. *Am J Epidemiol*. 2018;187(3):568–575. doi:10.1093/aje/kwx348
2. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64:402–406. doi:10.4097/kjae.2013.64.5.402
3. Wedderkopp N, Jespersen E, Franz C, et al. Study protocol. The Childhood Health, Activity, and Motor Performance School Study Denmark (The CHAMPS-study DK). *BMC Pediatr*. 2012;12:128. doi:10.1186/1471-2431-12-128
4. Little RJA, Rubin DB. *Statistical Analysis with Missing Data, Third Edition [Internet]*. 1st ed. John Wiley & Sons, Ltd; 2019. Available from https://onlinelibrary.wiley.com/doi/10.1002/9781119482260.
5. Little RJA, Rubin DB. *Complete-Case and Available-Case Analysis, Including Weighting Methods. Stat Anal Missing Data [Internet]*. John Wiley & Sons, Ltd; 2014:41–58. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119013563.ch3.
6. Newman DA. Missing Data: five Practical Guidelines. *Organ Res Methods*. 2014;17:372–411. doi:10.1177/1094428114548590
7. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157–166. doi:10.2147/CLEP.S129785
8. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. doi:10.1136/bmj.b2393
9. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59:1087–1091. doi:10.1016/j.jclinepi.2006.01.014
10. Rubin DB. Discussion on Multiple Imputation. *Int Stat Rev Rev Int Stat*. 2003;71:619–625. doi:10.1111/j.1751-5823.2003.tb00216.x
11. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am Stat*. 2007;61:79–90. doi:10.1198/000313007X172556
12. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med*. 2010;29:2920–2931. doi:10.1002/sim.3944
13. Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. *Int Stat Rev Rev Int Stat*. 2010;78:40–64. doi:10.1111/j.1751-5823.2010.00103.x
14. Rubin DB. The Bayesian Bootstrap. *Ann Stat*. 1981;9:130–134. doi:10.1214/aos/1176345338
15. Vaughan LK, Divers J, Padilla M, et al. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. *Comput Stat Data Anal*. 2009;53:1755–1766. doi:10.1016/j.csda.2008.02.032
16. Cattell RB, Jaspers J. A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivar Behav Res Monogr*. 1967;67–3:211.
17. Drechsler J. Multiple imputation in practice—a case study using a complex German establishment survey. *AStA Adv Stat Anal*. 2011;95:1–26. doi:10.1007/s10182-010-0136-z
18. Van Buuren S, Groothuis-Oudshoorn K. Multivariate imputation by chained equations. *J Stat Softw*. 2011;45:1–67.
19. R Core Team. *R: A Language and Environment for Statistical Computing [Internet]*. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: http://www.R-project.org. Accessed November 11, 2022.
20. Van Buuren S. *Flexible Imputation of Missing Data*. CRC Press; 2018.
21. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica*. 2012;22:276–282. doi:10.11613/BM.2012.031
22. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46:423–429. doi:10.1016/0895-4356(93)90018-V
23. Little RJ, Yosef M, Cain KC, Nan B, Harlow SD. A hot-deck multiple imputation procedure for gaps in longitudinal data on recurrent events. *Stat Med*. 2008;27:103–120. doi:10.1002/sim.2939
24. Wang C-N, Little R, Nan B, Harlow SD, Hot-Deck Multiple A. Imputation Procedure for Gaps in Longitudinal Recurrent Event Histories. *Biometrics*. 2011;67:1573–1582. doi:10.1111/j.1541-0420.2011.01558.x
25. Graham JW, Olchowski AE, Gilreath TD. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prev Sci*. 2007;8:206–213. doi:10.1007/s11121-007-0070-9