RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

# Challenge of missing data in observational studies: investigating cross-sectional imputation methods for assessing disease activity in axial spondyloarthritis

Stylianos Georgiadis ,[1] Marion Pons,[1] Simon Rasmussen,[1] Merete Lund Hetland ,[1,2] Louise Linde,[1] Daniela di Giuseppe,[3] Brigitte Michelsen,[1,4,5] Johan K Wallman,[6] Tor Olofsson ,[6] Jakub Zavada,[7,8] Bente Glintborg ,[1,2,9] Anne G Loft,[10,11] Catalin Codreanu,[12] Daniel Melim,[13] Diogo Almeida,[14] Sella Aarrestad Provan ,[4,15] Tore K Kvien ,[4,16] Vappu Rantalaiho,[17,18,19] Ritva Peltomaa,[20] Bjorn Gudbjornsson ,[12,21] Olafur Palsson,[6,12] Ovidiu Rotariu,[13] Ross MacDonald,[13] Ziga Rotar ,[14,22] Katja Perdan Pirkmajer,[14,22] Karin Lass,[23] Florenzo Iannone ,[24] Adrian Ciurea ,[25] Mikkel Østergaard ,[1,2] L M Ørnbjerg [1]

**To cite:** Georgiadis S, Pons M, Rasmussen S, *et al*. Challenge of missing data in observational studies: investigating cross-sectional imputation methods for assessing disease activity in axial spondyloarthritis. *RMD Open* 2025;**11**:e004844. doi:10.1136/rmdopen-2024-004844

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/rmdopen-2024-004844).

Received 6 August 2024
Accepted 20 January 2025



For numbered affiliations see end of article.

**Correspondence to**
Dr Stylianos Georgiadis;
stylianos.georgiadis@regionh.dk

## ABSTRACT

**Objectives** We aimed to compare various methods for imputing disease activity in longitudinally collected observational data of patients with axial spondyloarthritis (axSpA).

**Methods** We conducted a simulation study on data from 8583 axSpA patients from ten European registries. Disease activity was assessed by the Axial Spondyloarthritis Disease Activity Score (ASDAS) and the corresponding low disease activity (LDA; ASDAS<2.1) state at baseline, 6 and 12 months. We focused on cross-sectional methods which impute missing values of an individual at a particular time point based on the available information from other individuals at that time point. We applied nine single and five multiple imputation methods, covering mean, regression and hot deck methods. The performance of each imputation method was evaluated via relative bias and coverage of 95% confidence intervals for the mean ASDAS and the derived proportion of patients in LDA.

**Results** Hot deck imputation methods outperformed mean and regression methods, particularly when assessing LDA. Multiple imputation procedures provided better coverage than the corresponding single imputation ones. However, none of the evaluated methods produced unbiased estimates with adequate coverage across all time points, with performance for missing baseline data being worse than for missing follow-up data. Predictive mean and weighted predictive mean hot deck imputation procedures consistently provided results with low bias.

**Conclusions** This study contributes to the available methods for imputing disease activity in observational research. Hot deck imputation using predictive mean matching exhibited the highest robustness and is thus our suggested approach.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Missing data is a challenge in observational studies with data collected in routine care. Complete case analyses may result in bias and loss of power and imputation of missing data is advised.

⇒ Imputation methods using the available information from the same individual at other time points have been widely applied, but they can fail when a limited number of visits are available for an individual.

### WHAT THIS STUDY ADDS

⇒ The present work explores several cross-sectional imputation methods for use in observational studies, where missing values of an individual at a particular time point are imputed based on the available information from other individuals at that time point.

⇒ Our study shows that, among the evaluated cross-sectional imputation methods, imputation based on predictive mean or weighted predictive mean matching was the approach with the consistently best performance, and multiple imputation is strongly suggested.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The study informs the method selection process for imputing disease activity in observational research.

## INTRODUCTION

Axial spondyloarthritis (axSpA) is a chronic inflammatory disease that primarily involves the axial skeleton.[1] Various tools have been used to monitor axSpA, including inflammatory markers and patient-reported

outcomes.[2–4] The Axial Spondyloarthritis Disease Activity Score (ASDAS) is a widely used composite index that captures multiple important aspects of disease activity and is currently the preferred measure to assess disease activity in axSpA.[5–7] ASDAS is used to categorise patients into disease activity states with values of <1.3, <2.1 and >3.5 as the cut-offs between inactive disease, low, high and very high disease activity, respectively.[8 9]

Longitudinal observational studies on data from clinical registries offer the possibility to provide new evidence on the effectiveness and safety of treatments in patients in routine care.[10 11] However, patient-reported outcomes and disease activity measures are often missing in such studies, particularly during follow-up.[12 13]

A wide variety of methods have been developed to handle the challenges of missing data[14–16] and have been applied in observational studies.[17–19] A common approach is complete case analysis (CCA), which omits observations with missing values for any variable needed for analysis. However, this approach may result in bias in real-world settings, where data missingness depends on other observed variables. Another approach is substituting the missing variable of interest with one plausible value, that is, a single imputation. Single imputation is preferred over CCA in many settings because no observations are sacrificed, but this approach potentially produces biased results as well. A popular and more sophisticated alternative to deal with missing data is multiple imputation, where each missing value is replaced by several plausible values drawn from an imputation model. An overview of methods applied to observational data in rheumatology is presented by Lauper *et al*.[20]

Both single and multiple imputation methods can be divided into two categories: cross-sectional and longitudinal methods.[21] Cross-sectional methods use the available information from other individuals at a particular time point to impute missing values of an individual at that time point, while longitudinal methods use the available information from the same individual at other time points. While the performance of longitudinal imputation methods has been assessed in observational studies within rheumatology,[22] cross-sectional methods have not previously been investigated in this context. Cross-sectional methods may be more relevant in real-world data from multiple sources (eg, centres or registries), where the distribution and frequency of visits vary according to source and treatment, while the amount of missing data differs from one source to another. Using the information from other individuals, that is, cross-sectional methods, may thus be more relevant than using the information from the same individual longitudinally, as certain longitudinal methods may fail when a limited number of visits are available for an individual. Three different cross-sectional imputation procedures have been proposed: mean, linear regression and hot deck imputation.[17 21 23 24]

We aimed to compare a wide range of cross-sectional single and multiple imputation methods for assessing disease activity as measured by the mean ASDAS and the derived proportion of patients in low disease activity (LDA; ASDAS<2.1) at baseline, 6 and 12 months. Thus, we performed a simulation study on longitudinal observational data from a large multinational cohort of patients with axSpA.

## METHODS

### European Spondyloarthritis data

Analyses relied on pseudonymised prospectively collected data of patients registered with a diagnosis of axSpA in 10 registries participating in the European Spondyloarthritis (EuroSpA) research collaboration network; ATTRA (Czech Republic), biorx.si (Slovenia), BSRBR-AS (UK), DANBIO (Denmark), ICEBIO (Iceland), NOR-DMARD (Norway), Reuma.pt (Portugal), ROB-FIN (Finland), RRBR (Romania) and SRQ (Sweden). Commonalities and differences in set-up and clinical data collection across the registries participating in EuroSpA collaboration have previously been investigated by Linde *et al*.[25]

Patients eligible for inclusion in the present study were aged ≥18 years at the time of diagnosis, had a registered axSpA diagnosis and were initiating treatment with a tumour necrosis factor inhibitor (TNFi) or an interleukin 17A inhibitor (IL-17Ai) as their first biological disease-modifying anti-rheumatic drug between 1 January 2017 and 31 December 2021. Data at baseline and two follow-up visits (6 and 12 months) per treatment were collected within specified time windows; from 90 days before to 30 days after treatment initiation for baseline, from 90 days to 270 days after treatment initiation for the 6-month follow-up visits, and from 271 days to 450 days after treatment initiation for the 12-month follow-up visits.

### Cross-sectional imputation methods

The *variable of interest*, that is, ASDAS, is a continuous composite variable in which missing values occurred at a particular time point. In the current study, imputation of the ASDAS was assessed, not its individual components. Once ASDAS was imputed, the LDA state was determined from the imputed ASDAS value at the same time point.

Certain cross-sectional imputation methods applied in this study also make use of available information in *covariates*, that is, explanatory (complete) continuous or categorical variables at the same time point. Categorical covariates can be used to cross-classify the study population by creating *classes* that group individuals with similar characteristics. Continuous covariates may also be appropriately categorised for this purpose.

We applied both *deterministic* and *random* (or *stochastic*) imputation methods.[26] In deterministic imputation methods, there is no randomness involved in the selection of the imputed value, while in random imputation methods, the missing ASDAS values are substituted randomly from a set of plausible values.

**Table 1** Summary of the cross-sectional imputation methods

| Method | Imputations (n) | Short description | Group | Type | Parameter uncertainty* | Covariates |
|---|---|---|---|---|---|---|
| Unconditional M-SI | 1 | Mean of values observed in the total study population | Mean | Deterministic | – | – |
| Conditional M-SI | 1 | Mean of values observed in the corresponding class | Mean | Deterministic | – | Categorical |
| Deterministic LR-SI | 1 | Predicted value from a linear regression model | Regression | Deterministic | – | Any |
| Random LR-SI | 1 | Predicted value from a linear regression model with an error component term added | Regression | Random | Bootstrap[15] | Any |
| Unconditional HD-SI | 1 | Value of an individual drawn randomly from the total study population | Hot deck | Random | Approximate Bayesian bootstrap | – |
| Conditional HD-SI | 1 | Value of an individual drawn randomly from the corresponding class | Hot deck | Random | Approximate Bayesian bootstrap | Categorical |
| Predictive mean deterministic HD-SI | 1 | Value of the individual having the closest predictive mean | Hot deck | Deterministic | – | Any |
| Predictive mean random HD-SI | 1 | Value of an individual drawn randomly from a set containing 10 individuals with the closest predictive mean | Hot deck | Random | Bayesian[15] | Any |
| Weighted predictive mean random HD-SI | 1 | Value of an individual drawn randomly from the total study population with a probability that depends on the difference in predictive means | Hot deck | Random | Approximate Bayesian bootstrap[47] | Any |
| Random LR-MI | 10 | Multiple imputation with random linear regression method | Regression | Random | Bootstrap[15] | Any |
| Unconditional HD-MI | 10 | Multiple imputation with unconditional hot deck method | Hot deck | Random | Approximate Bayesian bootstrap | – |
| Conditional HD-MI | 10 | Multiple imputation with conditional hot deck method | Hot deck | Random | Approximate Bayesian bootstrap | Categorical |
| Predictive mean random HD-MI | 10 | Multiple imputation with predictive mean random hot deck method | Hot deck | Random | Bayesian[15] | Any |
| Weighted predictive mean random HD-MI | 10 | Multiple imputations with weighted predictive mean random hot deck method | Hot deck | Random | Approximate Bayesian bootstrap[47] | Any |

*Parameter uncertainty is incorporated into random imputation methods by a bootstrap, a Bayesian, or an approximate Bayesian bootstrap approach. In the bootstrap approach, parameters are estimated from a bootstrap sample drawn from the complete part of data, while, in the Bayesian approach, parameters are drawn from their posterior distribution, given the data.[15] Approximate Bayesian bootstrap is a two-step procedure which draws a bootstrap sample from the donor pool and then draws imputations randomly from the bootstrap sample.[48] Proper single and multiple imputation procedures are hence obtained for random methods, since they reflect all uncertainty, including uncertainty in predicting individual missing values given parameters and uncertainty in parameter estimation.[48]
ASAS, Assessment of SpondyloArthritis International Society; HD-MI, hot deck multiple imputation; HD-SI, hot deck single imputation; LR-MI, linear regression multiple imputation; LR-SI, linear regression single imputation; M-SI, mean single imputation.

Table 1 summarises the cross-sectional imputation procedures investigated in this study.

### Single imputation

We applied nine single imputation methods, divided into three groups, that is, mean, linear regression and hot deck imputation.

#### Mean imputation

Two alternative deterministic approaches to *mean single imputation (M-SI)*, where each missing ASDAS value is substituted by the mean of the values observed at that time point, are:

- *Unconditional M-SI*, where the mean is calculated based on the total study population.
- *Conditional M-SI*, where the mean is calculated within the appropriate class of the individual.

#### Linear regression imputation

*Linear regression single imputation (LR-SI)* can be seen as a generalisation of conditional M-SI, and it can be either deterministic or random:

- *Deterministic LR-SI*, where the missing ASDAS value is replaced by a value predicted from a linear regression of the complete cases of that variable on (complete) covariates at that time point.
- *Random LR-SI*, where randomness is assigned in the selection of the imputed value by adding an error component in the aforementioned linear regression method.

Values imputed by linear regression imputation were restricted to the range of observed values.

#### Hot deck imputation

*Hot deck single imputation (HD-SI)* uses a setup similar to conditional M-SI. The cross-sectional HD-SI involves a *recipient*, that is, the individual with a missing value in the variable of interest, and a *donor*, that is, another individual with similar characteristics to the recipient, and whose value is observed at the same time point. The missing value of the recipient is substituted by the observed value of a donor from a *donor pool*, that is, a set of potential donors. A hot deck imputation method can be either random, when the donor is selected randomly from a donor pool, or deterministic, when the donor is the closest donor to the recipient based on a certain matching metric. The same metric can be used to apply a random hot deck imputation by creating a donor pool for each recipient with a fixed number of candidate donors having the closest metric values, and then by randomly selecting a donor from the donor pool. In this study, we chose the predictive mean as the metric. Briefly, predictive mean matching calculates the predicted value of the variable of interest according to a specified imputation model and, for each recipient, the donor pool is formed from a fixed number of individuals with observed values that have predicted values closest to the predicted value for the recipient.[15] A review of hot deck imputation has been published by Andridge and Little.[27]

For the purpose of this study, we considered five hot deck imputation methods:

- *Unconditional HD-SI*, where a donor is sampled at random with replacement from the total study population.
- *Conditional HD-SI*, where a donor is sampled at random with replacement from the individuals in the recipient's class.
- *Predictive mean deterministic HD-SI*, where the donor having the minimal difference in predictive mean is selected.
- *Predictive mean random HD-SI*, where a donor is selected by a random draw from a pool containing the closest 10 candidate donors.[28]
- *Weighted predictive mean random HD-SI*, where a donor is drawn from the total study population with a probability that depends on their distance in the predictive metric from the recipient.[29]

For the conditional hot deck imputation method, we required a minimum of five recipients in each class.[27]

### Multiple imputation

Based on the random imputation methods described in the Single Imputation section, the following five cross-sectional multiple imputation methods were also applied; *random linear regression multiple imputation (random LR-MI)* and four random *hot deck multiple imputation (HD-MI): unconditional HD-MI*, *conditional HD-MI*, *predictive mean random HD-MI* and *weighted predictive mean random HD-MI*.

In multiple imputation, the whole process of a random single imputation method is repeated independently several times, and, in this way, various imputation values are calculated for every missing value. Briefly described, each missing value is replaced with $K \geq 2$ imputed values and $K$ complete datasets are created, where $K$ denotes the *number of imputations*. Estimates are derived in each imputed dataset and then inferences are combined across the $K$ imputed datasets, using pooling rules (namely Rubin's rules), to form one inference that reflects the uncertainty of the missing values. Details on the implementation of multiple imputations can be found in the literature.[14–16] In this study, the number of imputations was set to $K=10$.

### Complete case analysis

We also assessed CCA which disregards observations with missing ASDAS as a comparator method for imputation.

### Simulation framework

We conducted a simulation study to evaluate the cross-sectional imputation methods for assessing disease activity in axSpA.[30 31]

A simulated dataset of size $n_{obs}$ ($n_{obs}$=100, 200, 500 and 1000) was drawn at random from *complete case data*, that is, pooled registrations with complete data for ASDAS at a particular time point (baseline, 6 or 12 months after treatment initiation). To allow for comparison of methods over the increasing sequence of sample sizes,

the same starting seed in the random number generator for all sample sizes. The simulation process was independently repeated 1000 times by specifying the starting seed for each simulation repetition separated by at least the largest sample size, that is, 1000.[30] Missing values were introduced in each simulated dataset at a specific level of missingness ($\lambda$ = 10%, 20%, …, 90%), that is, the proportion of missing data to be created, and according to a chosen missing data mechanism, as described in the following section. Missing values were then imputed by applying the different cross-sectional imputation methods under study and a complete dataset with imputed missing values was eventually created.

The main analyses were carried out for a sample size of $n_{obs}$ = 1000 and at a level of missingness $\lambda$ = 60%, the same as in Mongin *et al*.[22] Simulations were conducted separately at each time point.

### Generation of missing data

Three different missing data mechanisms exist[14]: (a) missing completely at random (MCAR), when the probability of an observation being missing does not depend on any observed or unobserved (missing) variable; (b) missing at random (MAR), when the probability of an observation being missing depends only on the observed variables; and (c) missing not at random (MNAR), when the probability of an observation being missing additionally depends on one or more unobserved variables.

MCAR, MAR and MNAR data were generated independently at each time point. To generate MCAR data, registrations with missing data were selected from complete case data by random sampling without replacement. MAR data were induced in the simulated datasets by applying a propensity model.[32] In the model generating MAR values, we considered the following complete variables which may affect the missingness of ASDAS: registration year of visit, sex, age at registration year and gross domestic product (GDP) per capita at registration year (1000$). Data on GDP per capita at registration year for each country were retrieved from the International Monetary Fund's website.[33] To generate MNAR data, we applied a missing data rule according to which ASDAS had $\pi_1$ probability of being missing when ASDAS was below a threshold, otherwise ASDAS had $\pi_2$ probability of being missing. We hypothesised that registrations with lower disease activity had higher chances of being missing than higher disease activity and therefore $\pi_1 > \pi_2$. Registrations with MAR and MNAR values were sampled without replacement and with unequal probabilities via random systematic sampling.[34] As missing data are commonly neither MCAR nor MNAR,[35] results on MAR data were chosen to be primarily presented.

### Performance

For each parameter of the simulation framework (sample size, level of missingness, missing data mechanism and imputation method), bias, precision, accuracy and coverage of the applied imputation methods were assessed. The population parameter of interest was the expected value of ASDAS or the derived proportion of patients in LDA at a specific time point, while the true parameter values were determined from the complete case data. For each simulation repetition, the mean value of ASDAS and the proportion in ASDAS LDA, along with their standard errors, were obtained. Relative bias and coverage of 95% CIs constituted the primary endpoint of this study. By its definition, coverage should be approximately equal to the nominal coverage rate of 95% CIs, that is, 95%. Along with coverage, the average width of 95% CIs was evaluated. Since more narrow intervals translate into greater accuracy, if one method had a similar or higher coverage than another but yielded intervals that are substantially narrower, then it should be preferred.[36] Since different values of population parameters were used for different data-generating mechanisms, relative bias was preferred over absolute bias.[31] Bias, empirical SE and relative root mean squared error were also calculated and reported. Definitions of performance measures are shown in online supplemental table 1.

### Definition of classes and donor pools

Covariates are used for defining the classes and donor pools in certain cross-sectional imputation methods. In order to see a reduction in bias for the mean of ASDAS, the covariates must be associated with both the outcome and the binary variable indicating whether or not the outcome is missing.[27] We considered four complete covariates to create the classes and donor pools: registration year of visit, sex, age at registration year and GDP per capita at registration year (1000$).

### Additional analyses

To investigate the impact of the number of covariates in the imputation model on the performance of imputation methods, we considered a priori three additional covariates as predictors of ASDAS in the imputation model, that is, disease duration at registration year of visit (defined as the number of years since diagnosis), treatment type (TNFi or IL-17Ai) and human leucocyte antigen subtypes B27 (HLA-B27) positivity (yes vs no). For missing values at baseline, we also considered concomitant conventional synthetic disease-modifying anti-rheumatic drug (csDMARD) at treatment start (yes vs no). Among these four covariates, disease duration and concomitant csDMARD have missing values. However, multiple imputation as described previously assesses only the imputation of univariate missing data and therefore cannot accommodate incomplete covariates. We applied multiple imputation by chained equations (MICE) which allows for missing values in covariates used in the imputation model.[15] MICE iteratively imputes multivariate missing data on a variable-by-variable basis through a set of imputation models, one for each variable with missing values. Several iterations are thus needed to create a single imputed dataset. Different variable types, that is, continuous, binary, unordered and ordered categorical

variables, can be handled by MICE. The five cross-sectional multiple imputation methods, as described in a previous section, can be in principle incorporated into MICE for imputing ASDAS. However, unconditional hot deck method does not take covariates into account. Furthermore, conditional hot deck method requires solely categorical covariates and therefore cross-classification of the study population into classes based on a large number of covariates is unrealistic. Hence, we applied *random linear regression MICE (random LR-MICE), predictive mean random hot deck MICE (HD-MICE)* and *weighted predictive mean random HD-MICE* and compared them with corresponding multiple imputation methods *random LR-MI, predictive mean random HD-MI* and *weighted predictive mean random HD-MI* for a sample size of $n_{obs}$ = 1000 at a level of missingness $\lambda$ = 60%. Incomplete categorical covariates (HLA-B27 positivity and concomitant csDMARD) were imputed by logistic regression. The number of imputations was set to $K$=10 with 10 iterations.

### Impact of imputation methods on original data

The potential impact of the cross-sectional imputation procedures in the original data was investigated by comparing descriptive statistics for ASDAS and ASDAS LDA after imputing with CCA.

All analyses were conducted using R V.4.2.2 software.[37] Built-in functions of *mice* package[38] were used to implement deterministic and random linear regression imputation (*norm.predict* and *norm.boot*, respectively), predictive mean hot deck imputation (*pmm*) and weighted predictive mean hot deck imputation (*midastouch*). The code for implementing the cross-sectional single and multiple imputation methods applied in this study is provided as online supplemental file 1.

## RESULTS

### Cohort

Data from 8583 patients who had at least one available ASDAS registration at any time point, that is, at baseline, 6 or 12 months, were included in the analyses. The baseline characteristics for patients included in the analyses are shown in online supplemental table 2. Mean (SD) of ASDAS in complete case data at baseline, 6 months and 12 months were 3.7 (1.1), 2.0 (1.0) and 1.8 (0.9), respectively (table 2).

### Classes and donor pools

The levels of associations of registration year of visit, sex, age at registration year and GDP per capita at registration year (1000$) with ASDAS and missingness of ASDAS at each time point are summarised in online supplemental table 3. GDP per capita at registration year and sex were highly associated with both ASDAS and missingness of ASDAS at certain time points, leading to an increase in precision and a decrease in bias. Age was considered an important variable, although it would not affect bias. Thus, we used sex, age and GDP per capita to define classes (used in conditional mean and conditional hot deck imputation methods) and donor pools (used in predictive mean deterministic and random hot deck imputation methods). A cut-off of 45 years was chosen for age at registration year in accordance with ASAS criteria,[2] while a cut-off of 40 was chosen for GDP per capita at registration year (1000$), as there was a clear distinction of countries in EuroSpA data below and above 40, that is, GDP<40: Czech Republic, Portugal, Romania and Slovenia; and GDP≥40: Denmark, Finland, Iceland, Norway, Sweden and the UK.

**Table 2** Descriptive statistics of ASDAS components, ASDAS and ASDAS disease activity states in complete case data

| | Baseline | 6 months | 12 months |
|---|---|---|---|
| Number of registrations, n | 6753 | 6318 | 4389 |
| Patient global assessment, mean (SD) | 6.7 (2.4) | 3.2 (2.6) | 2.7 (2.4) |
| BASDAI Q2, mean (SD) | 6.9 (2.4) | 3.2 (2.7) | 2.9 (2.5) |
| BASDAI Q3, mean (SD) | 4.7 (3.2) | 2.2 (2.6) | 1.9 (2.4) |
| BASDAI Q6, mean (SD) | 6.1 (3.0) | 2.6 (2.6) | 2.2 (2.4) |
| CRP, mean (SD) | 18.2 (25.6) | 5.1 (9.5) | 5.2 (9.9) |
| ASDAS, mean (SD) | 3.7 (1.1) | 2.0 (1.0) | 1.8 (0.9) |
| ASDAS<2.1, n (%) | 545 (8.1%) | 3923 (62.1%) | 2994 (68.2%) |

Complete case data are defined as the pooled registrations with complete data for ASDAS at baseline, 6 or 12 months after treatment.
ASDAS consists of four questions reported by the patient (on back pain, peripheral pain/swelling, duration of morning stiffness and global disease activity) and CRP.[5] BASDAI questions are used to evaluate back pain (BASDAI Q2), peripheral pain/swelling (BASDAI Q3) and duration of morning stiffness (BASDAI Q6) in ASDAS formula, while the use of a CRP of 2 mg/L if CRP<2 mg/L is recommended,[49] ie, ASDAS=0.121×(BASDAI Q2)+0.110×PGA+0.073×(BASDAI Q3)+0.058×(BASDAI Q6)+0.579×ln(max(CRP,2)+1).
Phrasing of the individual BASDAI questions used in ASDAS formula:
BASDAI Q2. How would you describe the overall level of inflammatory neck, back or hip pain you have had?
BASDAI Q3. How would you describe the overall level of pain/swelling in joints other than neck, back or hips you have had?
BASDAI Q6. How long does your morning stiffness last from the time you wake up?
ASDAS, Axial Spondyloarthritis Disease Activity Score; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASDAI Q2, BASDAI question 2; BASDAI Q3, BASDAI question 3; BASDAI Q6, BASDAI question 6; CRP, C-reactive protein; PGA, Patient Global Assessment.

The same three variables were used as covariates in both linear regression imputation methods.

### MAR data

For each time point, 60% MAR data were introduced in each simulated dataset using a propensity model (online supplemental table 4). A summary of the results for the applied cross-sectional imputation methods in MAR data is presented in table 3.

### Disease activity

Coverage and average width for cross-sectional imputation methods assessing the mean ASDAS are depicted in figure 1 (upper panel). All performance measures are presented in online supplemental table 5.

For missing values at baseline, the only method with coverage close to 95% was weighted predictive mean random HD-MI (89.8%). This method also yielded the bias closest to zero (−0.20%), while the corresponding single imputation (weighted predictive mean random HD-SI) also estimated the mean ASDAS with a bias <1%. All predictive mean hot deck procedures (predictive mean deterministic HD-SI, predictive mean random HD-SI and predictive mean random HD-MI) had a bias <2%.

At 6 months, all multiple imputation methods (random LR-MI, unconditional HD-MI, predictive mean random HD-MI and weighted predictive mean random HD-MI) had comparable slight under-coverage, while the average width among them was smallest for unconditional HD-MI. Among them, weighted predictive mean random HD-MI had the lowest, but still similar bias (−0.52%). At 12 months, random LR-MI and weighted predictive mean random HD-MI had coverage above 90%, despite a slight over-estimation (3.49%) and under-estimation (−1.41%), respectively, of bias. Predictive mean random HD-MI had moderate under-coverage (88.1%) but gave close to unbiased results (−0.86%). For missing values at both follow-up time points, all single imputation methods had bias <5%. Linear regression and predictive mean hot deck procedures (deterministic LR-SI, random LR-SI, predictive mean deterministic HD-SI, predictive mean random HD-SI and weighted predictive mean random HD-SI) had a bias <1% at 6 months, whereas only the predictive mean hot deck ones (predictive mean deterministic HD-SI and predictive mean random HD-SI) had a bias <1% at 12 months.

Conditional HD-SI and conditional HD-MI failed to produce any results at any time point for MAR data, since the number of individuals in classes was not always sufficient in the simulated datasets. Conditional M-SI was available only for 6 months. CCA gave strongly biased results (>10%) and slightly biased results (<5%) for mean ASDAS at baseline and follow-up visits, respectively.

### Low disease activity

Performance for ASDAS LDA is shown in figure 1 (lower panel) and online supplemental table 6.

Regarding missing values at baseline, weighted predictive mean random HD-MI was the only method that estimated proportions of patients in LDA moderately well in terms of coverage (86.4%), while deterministic LR-SI and weighted predictive mean random HD-MI resulted in a bias <10%. When assessing ASDAS LDA at both follow-up visits, predictive mean random HD-MI, and weighted predictive mean random HD-MI overall performed fairly well in terms of bias (<5%) and coverage (>80%). Most of the single imputation methods had a bias <10%, while hot deck ones based on predictive mean matching (predictive mean deterministic HD-SI, predictive mean random HD-SI and weighted predictive mean random HD-SI) consistently resulted in a bias <5%.

We note that the true proportion of patients in ASDAS LDA was very low at baseline (8.1%, table 2) which may have influenced the performance assessment at this time point in comparison to follow-up visits (62.1% and 68.2% at 6 and 12 months, respectively). CCA yielded slightly to moderately biased results for missing data at follow-up visits, but not at baseline.

### Effect of level of missingness and sample size

Varying the level of missingness from 10% to 90% affected bias for mean ASDAS of single imputation methods differently (online supplemental figure 1). For methods that neglect available covariate information (unconditional M-SI and unconditional HD-SI) and CCA, bias increased almost linearly, as missingness increased. Methods that required only categorical covariates (conditional M-SI and conditional HD-SI) failed to produce results for levels of missingness above around 50% without a clear impact of the level of missingness on bias. For both linear regression methods (deterministic LR-SI and random LR-SI), bias increased as missingness increased. Level of missingness did not affect bias of predictive mean methods (predictive mean deterministic HD-SI, predictive mean random HD-SI and weighted predictive mean random HD-SI) for missingness up to around 60%, whereas bias increased substantially for missing data beyond that level, particularly at baseline. The impact of missingness on multiple imputation methods followed the same pattern as for the corresponding single ones. Bias was not heavily affected by sample size (results not shown).

Overall, coverage for all imputation procedures decreased as missingness increased (online supplemental figure 2). As expected, under-coverage due to bias tended to deteriorate, as sample size increased.[31]

### MCAR data

Regarding MCAR data, all imputation procedures and CCA gave unbiased, or close to unbiased, estimates of the mean ASDAS (online supplemental table 7), while multiple imputation methods yielded correct, or close to correct, coverage of 95% CIs with similar average widths (figure 2, upper panel). For ASDAS LDA, hot deck imputation procedures and CCA markedly outperformed mean and linear regression ones (online supplemental

**Table 3** Summary of the main findings for each imputation method in data missing at random

| Method | Performance measure* | ASDAS | | | ASDAS<2.1 | | |
|---|---|---|---|---|---|---|---|
| | | Baseline | 6 months | 12 months | Baseline | 6 months | 12 months |
| Unconditional M-SI | Bias | Strong over-estimation | Slight under-estimation | Slight under-estimation | Strong under-estimation | Strong over-estimation | Strong over-estimation |
| Conditional M-SI | Bias | – | Slight under-estimation | – | – | Moderate over-estimation | – |
| Deterministic LR-SI | Bias | Moderate under-estimation | Close to unbiased | Slight over-estimation | Moderate under-estimation | Moderate over-estimation | Moderate over-estimation |
| Random LR-SI | Bias | Moderate under-estimation | Close to unbiased | Slight over-estimation | Strong over-estimation | Moderate under-estimation | Moderate under-estimation |
| Unconditional HD-SI | Bias | Strong over-estimation | Slight under-estimation | Slight under-estimation | Strong under-estimation | Slight over-estimation | Moderate over-estimation |
| Conditional HD-SI | Bias | – | – | – | – | – | – |
| Predictive mean deterministic HD-SI | Bias | Slight under-estimation | Close to unbiased | Close to unbiased | Strong over-estimation | Slight over-estimation | Close to unbiased |
| Predictive mean random HD-SI | Bias | Slight under-estimation | Close to unbiased | Close to unbiased | Strong over-estimation | Slight over-estimation | Slight over-estimation |
| Weighted predictive mean random HD-SI | Bias | Close to unbiased | Close to unbiased | Slight under-estimation | Strong over-estimation | Close to unbiased | Slight over-estimation |
| Random LR-MI | Bias | Moderate under-estimation | Close to unbiased | Slight over-estimation | Strong over-estimation | Moderate under-estimation | Moderate under-estimation |
| | Coverage | Strong under-coverage | Slight under-coverage | Slight under-coverage | Strong under-coverage | Moderate under-coverage | Moderate under-coverage |
| Unconditional HD-MI | Bias | Strong over-estimation | Slight under-estimation | Slight under-estimation | Strong under-estimation | Slight over-estimation | Moderate over-estimation |
| | Coverage | Strong under-coverage | Slight under-coverage | Strong under-coverage | Strong under-coverage | Slight under-coverage | Strong under-coverage |
| Conditional HD-MI | Bias | – | – | – | – | – | – |
| | Coverage | – | – | – | – | – | – |
| Predictive mean random HD-MI | Bias | Slight under-estimation | Slight under-estimation | Close to unbiased | Strong over-estimation | Slight over-estimation | Slight over-estimation |
| | Coverage | Strong under-coverage | Slight under-coverage | Moderate under-coverage | Strong under-coverage | Slight under-coverage | Moderate under-coverage |

Continued

**Table 3** Continued

| Method | Performance measure* | ASDAS | | | ASDAS<2.1 | | |
|---|---|---|---|---|---|---|---|
| | | Baseline | 6 months | 12 months | Baseline | 6 months | 12 months |
| Weighted predictive mean random HD-MI | Bias | Close to unbiased | Close to unbiased | Slight under-estimation | Moderate over-estimation | Slight over-estimation | Slight over-estimation |
| | Coverage | Moderate under-coverage | Slight under-coverage | Slight under-coverage | Moderate under-coverage | Slight under-coverage | Moderate under-coverage |

'–' indicates that results were not produced due to insufficient number of individuals in classes in the simulated datasets.
Assessment of bias: close to unbiased (<1%), slight bias (1%–5%), moderate bias (5%–10%) and strong bias (>10%). Assessment of coverage: correct coverage (Monte Carlo 95% CIs of coverage include 95%), slight under-coverage (90%–95%), moderate under-coverage (80%–90%), strong under-coverage (<80%) and over-coverage (95%–100%). Monte Carlo 95% CIs of coverage were calculated based on corresponding 1.96 Monte Carlo SEs.
*Strong under-coverage was observed for all single imputation methods.
ASDAS, Axial Spondyloarthritis Disease Activity Score; HD-MI, hot deck multiple imputation; HD-SI, hot deck single imputation; LR-MI, linear regression multiple imputation; LR-SI, linear regression single imputation; M-SI, mean single imputation.

table 8). Hot deck multiple imputation methods gave correct, or close to correct, coverage of 95% CIs (figure 2, lower panel). Overall, conditional HD-SI and conditional HD-MI gave similar results as the corresponding hot deck imputation methods.

### MNAR data

To generate MNAR data at baseline, we assumed that ASDAS had a probability $\pi_1 = 0.2$ of being missing when ASDAS<3.5 (ie, the cut-off between high and very high disease activity), otherwise ASDAS had a probability $\pi_2 = 0.1$ of being missing. At 6- and 12-month follow-up visits, we assumed the same probabilities but at a threshold of ASDAS<2.1 (ie, the cut-offs between low and high disease activity). All imputation procedures and CCA failed in terms of bias and coverage of 95% CIs for both mean ASDAS and ASDAS LDA (online supplemental tables 9 and 10).

### Additional analyses

The three cross-sectional imputation methods using MICE (random LR-MICE, predictive mean random HD-MICE and weighted predictive mean random HD-MICE) provided overall results with similar bias with the corresponding multiple imputation methods (random LR-MI, predictive mean random HD-MI and weighted predictive mean random HD-MI) under MCAR, MAR or MNAR assumptions (online supplemental tables 11 and 12). For MAR or MCAR data, random LR-MICE returned lower coverage of 95% CIs than random LR-MI, while predictive mean random HD-MICE and weighted predictive mean random HD-MICE mostly improved coverage slightly (online supplementary tables 11 and 12; online supplementary figures 3 and 4).

### Impact of imputation methods on original data

Mean ASDAS and proportion of patients with ASDAS<2.1 in data from 11 011 patients with at least one available visit at any time point, that is, at baseline, 6 or 12 months, were calculated according to all cross-sectional imputation procedures and CCA (online supplemental table 13). The covariates included in the imputation models were the same as in the simulation analyses. The proportions of missing data in ASDAS were 29%, 24% and 21% at baseline, 6 months and 12 months, respectively. Multiple imputation procedures provided estimates very close to the corresponding single imputation ones. As compared with CCA, hot deck imputation procedures using predictive mean matching gave a somewhat lower estimate of ASDAS at baseline (weighted predictive mean HD-MI: 3.54 vs CCA: 3.68), but similar estimates at follow-up visits, while slight discrepancies were observed for the proportions of patients with ASDAS<2.1 at baseline and 6 months.

## DISCUSSION

This simulation study evaluated the performance of a wide range of cross-sectional single and multiple

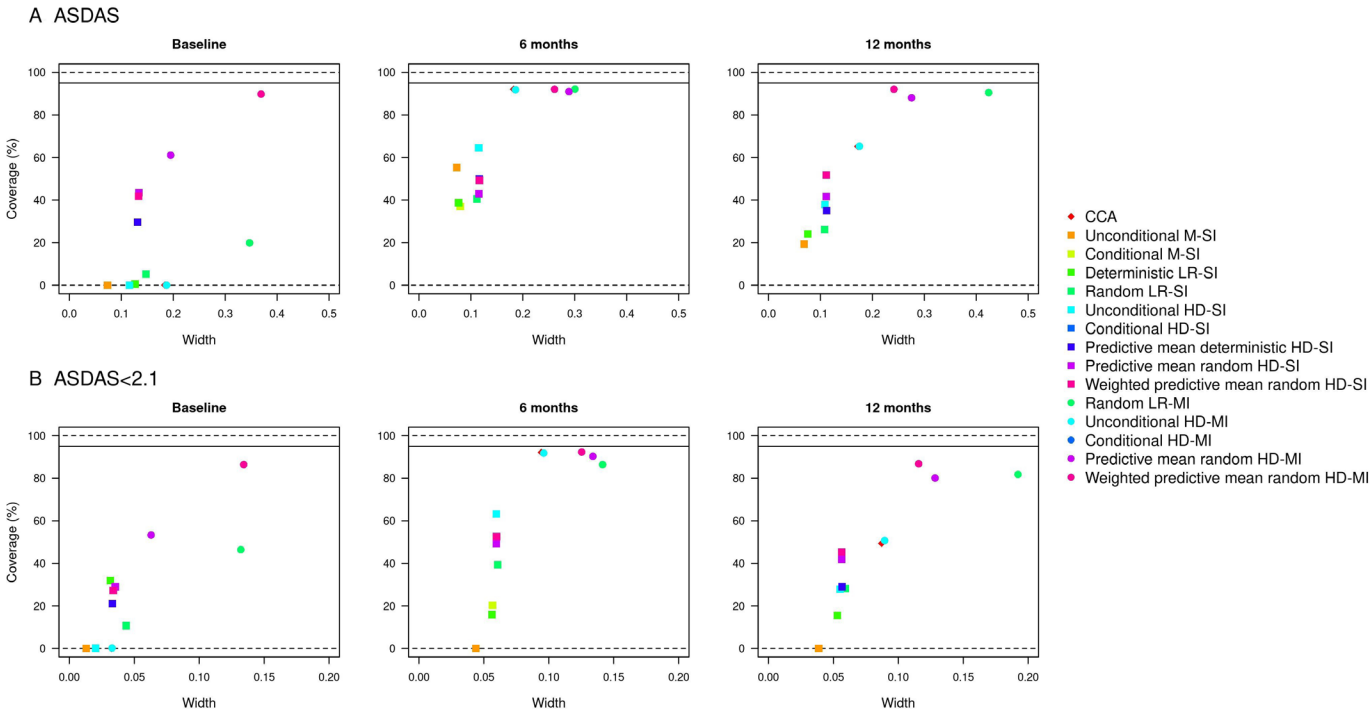


**Figure 1** Coverage and average width of 95% CIs in data of sample size 1000, where 60% of data were missing at random for (A) ASDAS, and (B) ASDAS<2.1. Coverage of 95% CIs is defined as the proportion of times that the 95% CI contains the true parameter value. Average width of 95% CIs is defined as the difference of mean lower and upper limits of the 95% CIs. Symbols □ and ○ indicate a single and a multiple imputation method, respectively. The same colour is used for a certain method in both a single and a multiple imputation setting. Solid black horizontal lines represent the nominal coverage rate (ie, 95%). Dashed black horizontal lines represent the range of coverage (ie, 0% and 100%). ASDAS, Axial Spondyloarthritis Disease Activity Score; CCA, complete case analysis; HD-MI, hot deck multiple imputation; HD-SI, hot deck single imputation; LR-MI, linear regression multiple imputation; LR-SI, linear regression single imputation; M-SI, mean single imputation.

imputation methods for assessing disease activity in observational longitudinal studies, using real-life axSpA data from 10 different country-specific registers from the EuroSpA collaboration. Our results demonstrated that hot deck methods outperformed mean and linear regression methods, particularly when assessing LDA. However, none of the evaluated methods produced unbiased estimates with correct coverage of 95% CIs across all time points, with performance for missing baseline data being worse than for missing follow-up data. Weighted predictive mean hot deck multiple imputations was the approach with the consistently best performance in terms of bias and coverage of 95% CIs, while corresponding HD-SI methods only provided results with low bias.

Two of the main strengths of hot deck methods over mean and regression approaches are that they impute real values and that they can also handle other types of variables of interest than continuous ones. Additionally, hot deck methods avoid assumptions of linear regression and restrict imputed values to the range of possible values. However, the performance of hot deck methods may be affected by the availability of close donors which depends on the sample size and the level of missingness. Various imputation methods incorporate covariate information at a different level, for example, continuous or categorical (table 1). On the one hand, unconditional mean and unconditional hot deck imputation methods neglect available covariate information. On the other hand, conditional mean and conditional hot deck imputation methods require categorical variables to create classes which may not be feasible when many covariates are considered, particularly for high levels of missingness. Moreover, choosing appropriate cut-off values to categorise continuous covariates may be challenging. Regarding the type of imputation methods, random methods implemented in this study incorporate sampling variability into the parameters by a bootstrap, a Bayesian or an approximate Bayesian bootstrap approach (table 1), on top of the randomness that is involved in the selection of the imputed values. Taking all these points into account, predictive mean and weighted predictive mean random hot deck imputation methods are the most alluring among the single imputation approaches under study, as also verified by our results.

Single imputation allows standard complete data analyses to be applied. In our analyses, single imputation procedures resulted in bias comparable to the corresponding multiple imputation ones. Nevertheless, a key problem with single imputation methods is that they treat imputed values as observed ones. Inferences based on the single imputed dataset do not account for imputation uncertainty due to the missing information, thus yielding SEs that are too small.[14] As was also observed in this study, small SEs lead to under-coverage of 95% CIs.[39] Multiple

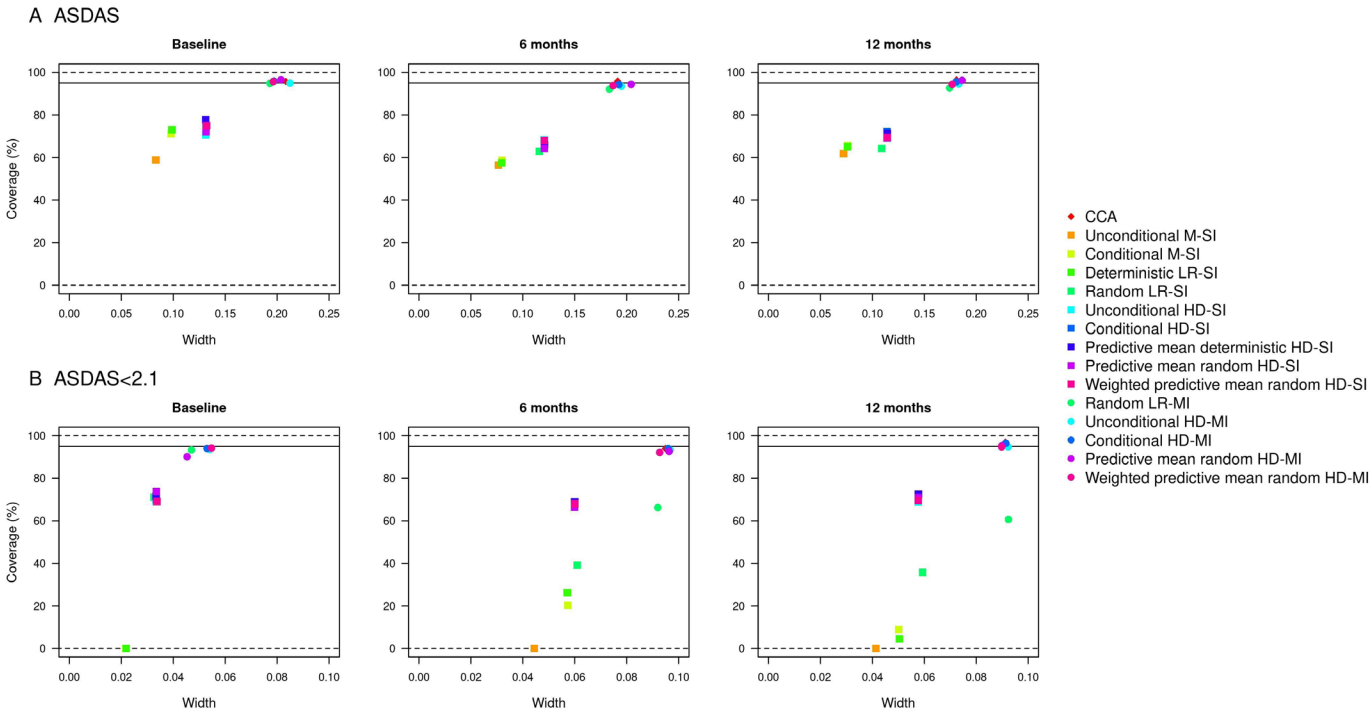


**Figure 2** Coverage and average width of 95% CIs in data of sample size 1000, where 60% of data were missing completely at random for (A) ASDAS, and (B) ASDAS<2.1. Coverage of 95% CIs is defined as the proportion of times that the 95% CI contains the true parameter value. Average width of 95% CIs is defined as the difference of mean lower and upper limits of the 95% CIs. Symbols □ and ○ indicate a single and a multiple imputation method, respectively. The same colour is used for a certain method in both a single and a multiple imputation setting. Solid black horizontal lines represent the nominal coverage rate (ie, 95%). Dashed black horizontal lines represent the range of coverage (ie, 0% and 100%). ASDAS, Axial Spondyloarthritis Disease Activity Score; CCA, complete case analysis; HD-MI, hot deck multiple imputation; HD-SI, hot deck single imputation; LR-MI, linear regression multiple imputation; LR-SI, linear regression single imputation; M-SI, mean single imputation.

imputation incorporates the uncertainty associated with missing data and therefore rectifies this disadvantage.

In additional analyses, we implemented three cross-sectional imputation methods via MICE, a very popular strategy for imputing multivariate missing data which has been widely applied in clinical research.[40–42] Covariates used in the imputation model can have missing values which is an advantage of MICE over the multiple imputation approach mainly investigated in our study. Therefore demographics, clinical measures and patient-reported outcomes could be used as covariates in the model to impute a disease activity measure. However, including additional covariates in the imputation model did not provide drastically better performance of the relevant imputation procedures. Despite that we considered the additional variables as clinically relevant, they may not be strong predictors of ASDAS. Besides, a higher number of iterations would potentially improve the performance of imputation methods in the MICE setting. We mention that the two hot deck imputation methods with predictive mean matching again outperformed linear regression imputation. In practice, when imputing components of composite disease activity scores or multiple disease activity scores at the same time, one can consider the flexible MICE approach in a cross-sectional setting with a predictive mean or weighted predictive mean random hot deck method.

One major pitfall when implementing any imputation approach is the misspecification of the imputation model.[43] If there are strong associations in the data that are not modelled, performance can become poor.[28 43] It is highly recommended to include all variables that are used in the analysis model and the predictors of the incomplete variable in the imputation model.[44] As also discussed in Mongin *et al*,[22] even though the covariates used in the main analyses were also used in the mechanism generating MAR data, disease activity could not be adequately imputed under this missing data assumption. Misspecification of the imputation model may have caused substantial bias in these analyses, which in turn resulted in poor coverage of 95% CIs.[31] Different relations in the data and different mechanisms generating MAR data may also explain discrepancies in performance results across assessment time points. Nevertheless, imputation based on predictive mean matching has been found to mitigate imputation model misspecification under MAR assumption.[28 43] We also note that substantial biases may occur under MAR and MNAR, because the complete cases are often unrepresentative of the population.[16] Moreover, for MNAR data, a distribution

for the missingness must be explicitly specified.[16] The aforementioned observations can explain why all imputation methods were strongly biased in the simple MNAR scenario presented in this study.

In the imputation model used in this study, ASDAS LDA was derived directly from ASDAS, adopting an 'impute, then transform' (or 'passive') approach.[44–46] This approach preserves the derivation relation between ASDAS and ASDAS disease activity state, but the derived variable is not part of the imputation model, which can lead to bias. This fact may explain the poor performance of certain methods when assessing the proportion of patients in LDA, even under the MCAR assumption. An alternative would be to include the derived variable in the imputation model and impute it directly like any other variable, known as 'transform, then impute'[45] or 'just another variable'.[44] This approach incorporates all variables into the imputation model, but it can lead to inconsistencies between ASDAS and ASDAS LDA. Literature addressing 'impute, then transform' and 'transform, then impute' has been contradictory. However, the 'transform, then impute' approach is recommended over 'impute, then transform',[45] although an 'impute, then transform' approach using predictive mean matching gave results comparable with those from 'transform, then impute'.[44] Another study showed that the choice of strategy for imputing a binary outcome variable depended on the level of missingness.[46] In analyses where only composite disease activity scores, disease activity states or other derived response variables are of interest, one may consider applying the 'transform, then impute' strategy.

All findings point to using an imputation procedure based on predictive mean matching. These approaches showed overall superior performance over mean and linear regression ones across endpoints and time points and a robustness to potential misspecification of the imputation model. Nevertheless, a predictive mean hot deck procedure, similarly to any hot deck procedures, requires good matches of donors to the recipient that reflect available covariate information.[27] Finding good matches is more likely in large than in small samples but also depends on the level of missingness. In most cases, hot deck imputation with weighted predictive mean matching performed slightly better than predictive mean matching. The weighted predictive mean matching approach considers all individuals in the donor pool, while the donor pool in predictive mean matching is restricted to a fixed number of candidate donors, that is, 10 in this study. However, a higher number of candidate donors might be needed in larger datasets.[28]

Both longitudinal and cross-sectional imputation methods can be used to replace missing data in longitudinal studies. Longitudinal imputation methods are often preferred over cross-sectional ones.[17 21] Comparing the cross-sectional single imputation methods of this study with the corresponding longitudinal ones presented earlier by Mongin *et al*,[22] such a preference does not appear to be justified. For the same level of missingness of 60%, certain cross-sectional imputation methods seem to be superior to the longitudinal imputation methods in terms of relative bias, under both MCAR and MAR assumptions. Investigators can evaluate whether imputing missing data using either the available longitudinal data of the same individual or the available information from other individuals at a particular time point is preferred according to the specific study setting.

This study highlights the potential of cross-sectional imputation methods for disease activity in observational studies in axSpA and may contribute to their implementation in practice. Our simulation results demonstrate the robustness of hot deck imputation procedures using predictive mean matching, which is thus the suggested approach of those evaluated.

**Author affiliations**
[1]Copenhagen Center for Arthritis Research (COPECARE), Center for Rheumatology and Spine Diseases, Center of Head and Orthopaedics, Rigshospitalet, Glostrup, Denmark
[2]Department of Clinical Medicine, University of Copenhagen, Kobenhavn, Denmark
[3]Clinical Epidemiology Division, Department of Medicine Solna, Karolinska Institutet, Solna, Sweden
[4]Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway
[5]Research Unit, Sørlandet Hospital, Kristiansand, Norway
[6]Department of Clinical Sciences Lund, Rheumatology, Skåne University Hospital, Lund University, Lund, Sweden
[7]Institute of Rheumatology, Prague, Czech Republic
[8]Department of Rheumatology, First Faculty of Medicine, Charles University, Praha, Czech Republic
[9]DANBIO registry, Rigshospitalet, Glostrup, Denmark
[10]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[11]Department of Rheumatology, Aarhus University Hospital, Aarhus, Denmark
[12]Faculty of Medicine, University of Iceland, Reykjavik, Iceland
[13]Aberdeen Centre for Arthritis and Musculoskeletal Health (Epidemiology Group), University of Aberdeen, Aberdeen, UK
[14]Department of Rheumatology, University Medical Centre Ljubljana, Ljubljana, Slovenia
[15]Public Health Section, Inland Norway University of Applied Sciences, Elverum, Norway
[16]Faculty of Medicine, University of Oslo, Oslo, Norway
[17]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[18]Centre for Rheumatic Diseases, Tampere University Hospital, Tampere, Finland
[19]Department of Medicine, Kanta-Häme Central Hospital, Hämeenlinna, Finland
[20]Rheumatology, Inflammation Center, Helsinki University Central Hospital, Helsinki, Finland
[21]Centre for Rheumatology Research, Landspitali University Hospital, Reykjavik, Iceland
[22]Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
[23]Department of Rheumatology, East Tallinn Central Hospital, Tallinn, Estonia
[24]Rheumatology Unit, University of Bari, Bari, Italy
[25]Department of Rheumatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

**X** Mikkel Østergaard @None

**Acknowledgements** The EuroSpA collaboration has been supported by Novartis Pharma AG since 2017 and UCB Biopharma SRL since 2022. This EuroSpA study was financially supported by UCB. No financial sponsors had any influence on the data collection, statistical analyses, manuscript preparation or decision to submit.

**Contributors** SG conceived of the original idea. SG and LMO designed the analyses. SG performed the simulations and analysed the data. SG drafted the manuscript with input from MP, SR, MLH, LL, DdG, BM, MØ and LMO. LMO supervised the project. All authors revised critically the manuscript and agreed with

its content. For further information, see CRediT statement in online supplemental file 2. SG is responsible for the overall content as the guarantor.

**Funding** This work was supported by UCB. UCB had no influence on the data collection, statistical analyses, manuscript preparation or decision to submit the manuscript.

**Competing interests** SG: Novartis, UCB; MP: Novartis, UCB; SR: Novartis, UCB; MLH: Abbvie, Biogen, BMS, Celltrion, Eli Lilly, Janssen Biologics B.V, Lundbeck Fonden, Medac, MSD, Novartis, Nordforsk, Pfizer, Roche, Samsung Biopies, Sandoz; LL: Novartis, UCB; DdG: none; BM: Novartis; JKW: AbbVie, Amgen, Eli Lilly, Novartis, Pfizer; TO: MSD, UCB; JZ: Abbvie, AstraZeneca, Egis, Elli Lilly, Novartis, Sandoz, Sanofi, Sobi, UCB; BG: Abbvie, BMS, Pfizer, Sandoz; AGL: AbbVie, Janssen, Eli Lilly, Novartis, Pfizer, UCB; CC: AbbVie, Amgen, AstraZeneca, Boehringer Ingelheim, Ewopharma, Eli Lilly, Novartis, Pfizer, Sandoz, Sobi; DM: none; DA: none; SAP: Boehringer Ingelheim; TKK: AbbVie, Amgen, BMS, Celltrion, Galapagos, Gilead, Grünenthal, Novartis, Pfizer, Sandoz, UCB; VR: Abbvie, BMS, Lilly, Novartis, Viatris; RP: Abbvie, Boehringer Ingelheim, Celltrion, Eli Lilly, Fresenius, Galapacos. Janssen, UCB; BG: none; OP: none; OR: none; RMD: none; ZR: Abbvie, Amgen, AstraZeneca, Biogen, Eli Lilly, Janssen, Lek, Medis, MSD, Novartis, Pfizer, Sanofi, Sobi, Swixx BioPharma; KPP: Abbvie, Boehringer Ingelheim, Eli Lilly, Medis, MSD, Lek, Novartis, Pfizer; KL: Abbvie, Johnson and Johnson, Novartis, Pfizer; FI: Abbvie, Amgen, AstraZeneca, BMS, Eli Lilly, Galapagos, Janssen, MSD, Novartis, Pfizer, UCB; AC: none; MØ: Abbvie, Amgen, BMS, Boehringer Ingelheim, Celgene, Eli Lilly, Galapagos, Gilead, Hospira, Janssen, MEDAC, Merck, Novartis, Novo, Orion, Pfizer, Regeneron, Roche, Sandoz, Sanofi, UCB; LMO: Novartis, UCB.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. Data may be obtained from a third party and are not publicly available. The data in this article was collected in the individual registries and made available for secondary use through the EuroSpA Research Collaboration Network https://eurospa.eu/# registries. Relevant patient level data may be made available on reasonable request to the corresponding author but will require approval from all contributing registries.



**ORCID iDs**
Stylianos Georgiadis http://orcid.org/0000-0003-3485-9457
Merete Lund Hetland http://orcid.org/0000-0003-4229-6818
Tor Olofsson http://orcid.org/0000-0002-9919-4487
Bente Glintborg http://orcid.org/0000-0002-8931-8482
Sella Aarrestad Provan http://orcid.org/0000-0001-5442-902X
Tore K Kvien http://orcid.org/0000-0002-8441-3093
Bjorn Gudbjornsson http://orcid.org/0000-0003-4631-6505
Ziga Rotar http://orcid.org/0000-0002-9323-9189
Florenzo Iannone http://orcid.org/0000-0003-0474-5344
Adrian Ciurea http://orcid.org/0000-0002-7870-7132
Mikkel Østergaard http://orcid.org/0000-0003-3690-467X
L M Ørnbjerg http://orcid.org/0000-0002-7832-6831

## REFERENCES

1 Sieper J, Poddubnyy D. Axial spondyloarthritis. *The Lancet* 2017;390:73–84.
2 Sieper J, Rudwaleit M, Baraliakos X, *et al*. The Assessment of SpondyloArthritis international Society (ASAS) handbook: a guide to assess spondyloarthritis. *Ann Rheum Dis* 2009;68 Suppl 2:ii1–44.
3 Landewé R, van Tubergen A. Clinical Tools to Assess and Monitor Spondyloarthritis. *Curr Rheumatol Rep* 2015;17:47.
4 Navarro-Compán V, Sepriano A, El-Zorkany B, *et al*. Axial spondyloarthritis. *Ann Rheum Dis* 2021;80:1511–21.
5 Lukas C, Landewé R, Sieper J, *et al*. Development of an ASAS-endorsed disease activity score (ASDAS) in patients with ankylosing spondylitis. *Ann Rheum Dis* 2009;68:18–24.
6 Ramiro S, Nikiphorou E, Sepriano A, *et al*. ASAS-EULAR recommendations for the management of axial spondyloarthritis: 2022 update. *Ann Rheum Dis* 2023;82:19–34.
7 van der Heijde D, Molto A, Ramiro S, *et al*. Goodbye to the term 'ankylosing spondylitis', hello 'axial spondyloarthritis': time to embrace the ASAS-defined nomenclature. *Ann Rheum Dis* 2024;83:547–9.
8 Machado P, Landewé R, Lie E, *et al*. Ankylosing Spondylitis Disease Activity Score (ASDAS): defining cut-off values for disease activity states and improvement scores. *Ann Rheum Dis* 2011;70:47–53.
9 Machado PM, Landewé R, Heijde D van der, *et al*. Ankylosing Spondylitis Disease Activity Score (ASDAS): 2018 update of the nomenclature for disease activity states. *Ann Rheum Dis* 2018;77:1539–40.
10 Stürmer T, Wang T, Golightly YM, *et al*. Methodological considerations when analysing and interpreting real-world data. *Rheumatology (Oxford)* 2020;59:14–25.
11 Courvoisier DS, Lauper K, Kedra J, *et al*. EULAR points to consider when analysing and reporting comparative effectiveness research using observational data in rheumatology. *Ann Rheum Dis* 2022;81:780–5.
12 Molto A, Tezenas du Montcel S, Wendling D, *et al*. Disease activity trajectories in early axial spondyloarthritis: results from the DESIR cohort. *Ann Rheum Dis* 2017;76:1036–41.
13 Christiansen SN, Ørnbjerg LM, Rasmussen SH, *et al*. European bio-naïve spondyloarthritis patients initiating TNF inhibitor: time trends in baseline characteristics, treatment retention and response. *Rheumatology (Sunnyvale)* 2022;61:3799–807.
14 Little RJA, Rubin DB. Statistical Analysis with Missing Data3rd ed. Hoboken, NJ: John Wiley & Sons, 2020.
15 Buuren S. Flexible Imputation of Missing Data2nd ed. Boca Raton, FL: Chapman and Hall/CRC Press, 2018.
16 Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods* 2002;7:147–77.
17 Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. *J Clin Epidemiol* 2003;56:968–76.
18 Pedersen AB, Mikkelsen EM, Cronin-Fenton D, *et al*. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol* 2017;9:157–66.
19 Harel O, Mitchell EM, Perkins NJ, *et al*. Multiple Imputation for Incomplete Data in Epidemiologic Studies. *Am J Epidemiol* 2018;187:576–84.
20 Lauper K, Kedra J, de Wit M, *et al*. Analysing and reporting of observational data: a systematic review informing the EULAR points to consider when analysing and reporting comparative effectiveness research with observational data in rheumatology. *RMD Open* 2021;7:1–9.
21 Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. *J Clin Epidemiol* 2002;55:329–37.
22 Mongin D, Lauper K, Turesson C, *et al*. Imputing missing data of function and disease activity in rheumatoid arthritis registers: what is the best technique? *RMD Open* 2019;5:e000994.
23 Pérez A, Dennis RJ, Gil JFA, *et al*. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. *Stat Med* 2002;21:3885–96.
24 Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *Am J Epidemiol* 2004;160:34–45.
25 Linde L, Ørnbjerg LM, Rasmussen SH, *et al*. Commonalities and differences in set-up and data collection across European spondyloarthritis registries - results from the EuroSpA collaboration. *Arthritis Res Ther* 2023;25:205.
26 Nordholt ES. Imputation: Methods, Simulation Experiments and Practical Examples. *Int Statistical Rev* 1998;66:157–80.
27 Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. *Int Stat Rev* 2010;78:40–64.
28 Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* 2014;14:75.
29 Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Stat Med* 2008;27:83–102.
30 Burton A, Altman DG, Royston P, *et al*. The design of simulation studies in medical statistics. *Stat Med* 2006;25:4279–92.
31 Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med* 2019;38:2074–102.

32 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
33 International Monetary Fund [Internet]. GDP per capita, current prices, Available: https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD
34 Tillé Y. Sampling and Estimation from Finite Populations. Hoboken, NJ: John Wiley & Sons, 2020.
35 Donders ART, van der Heijden GJMG, Stijnen T, *et al*. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087–91.
36 Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001;6:330–51.
37 R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2024.
38 Buuren S, mice G-O. Multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:1–67.
39 White IR, Pham TM, Quartagno M, *et al*. How to check a simulation study. *Int J Epidemiol* 2024;53:1–7.
40 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.
41 Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol* 2015;15:30:30:.
42 Austin PC, White IR, Lee DS, *et al*. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can J Cardiol* 2021;37:1322–31.
43 Curnow E, Carpenter JR, Heron JE, *et al*. Multiple imputation of missing data under missing at random: compatible imputation models are not sufficient to avoid bias if they are mis-specified. *J Clin Epidemiol* 2023;160:100–9.
44 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011;30:377–99.
45 von Hippel PT. 8. How to Impute Interactions, Squares, and other Transformed Variables. *Sociol Methodol* 2009;39:265–91.
46 Floden L, Bell ML. Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study. *BMC Med Res Methodol* 2019;19:161.
47 Gaffert P, Meinfelder F, Bosch V. Towards an mi-proper predictive mean matching. Bamberg, Germany, 2016.
48 Rässler S, Rubin DB, Zell ER. Imputation. *WIREs Computational Stats* 2013;5:20–9.
49 Machado P, Navarro-Compán V, Landewé R, *et al*. Calculating the ankylosing spondylitis disease activity score if the conventional c-reactive protein level is below the limit of detection or if high-sensitivity c-reactive protein is used: an analysis in the DESIR cohort. *Arthritis Rheumatol* 2015;67:408–13.