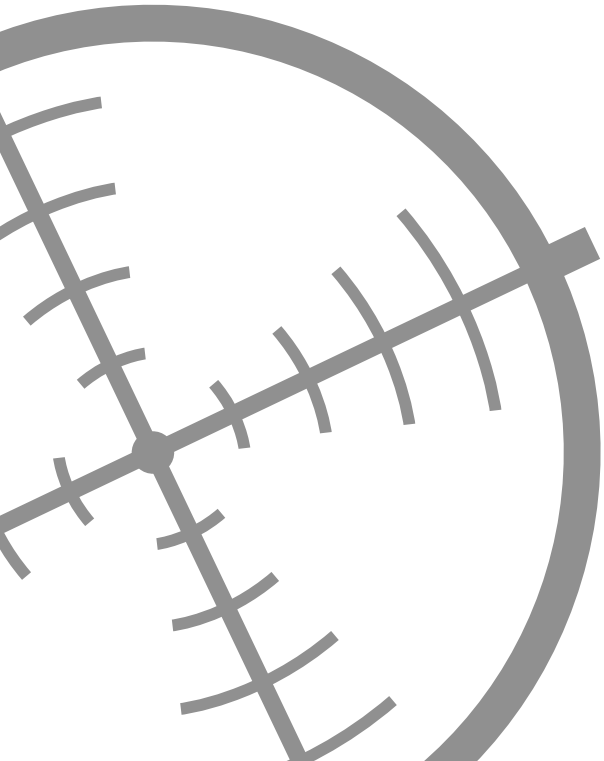
4

Linear extrapolation of missing radiographic change scores in clinical trials does not spuriously overestimate group radiographic changes in rheumatoid arthritis

Iris Markusse
Robert Landewé
Ron Wolterbeek
Meilien Ho
Martin Jenkins
Désirée van der Heijde



Submitted

ABSTRACT

Objective

To assess linear extrapolation (LE) and last observation carried forward (LOCF) as imputation methods for radiographic change in patients with rheumatoid arthritis (RA).

Methods

The OSKIRA-1 trial enrolled 918 patients with active RA to study the efficacy of fostamatinib. Radiographs were scheduled in all patients at baseline and week 12, regardless of early escape, and at week 24 and 52 in patients who remained on study. Complete radiographic data for 24 week follow-up were available for 623 patients, and were assessed according to the Sharp/ van der Heijde score (SHS). From this complete set of data a random selection of 10% missingness was generated. This was done 1000 times and for each replicate the missing radiographic change at week 24 was imputed either by LE, or by LOCF. The mean of the mean and standard deviation (SD) across the 1000 replications was calculated. A similar approach was iterated for different proportions of missingness.

Results

Mean (SD) observed SHS change from baseline to week 24 was 0.36 (2.39). With LE, mean (SD) change was estimated as 0.36 (2.65), 0.35 (2.88), 0.35 (3.17), 0.34 (3.57), 0.32 (4.45) with 10/20/30/50/90% missingness, respectively. With LOCF, mean (SD) change was estimated as 0.34 (2.39), 0.32 (2.38), 0.30 (2.37), 0.26 (2.36), 0.18 (2.34) with 10/20/30/50/90% missingness, respectively.

Conclusions

Linear extrapolation leads to stable estimates of progression at the group level but increasing variability by increasing proportions of missingness. In contrast, LOCF imputation systematically underestimates mean progression by increasing proportions of missingness, with artificially reduced variability estimates.

INTRODUCTION

In clinical trials involving patients with rheumatoid arthritis (RA), joint damage progression assessed on conventional radiographs of hands and feet is traditionally an important outcome measure.[1–3] Inevitably, one has to deal with missing radiographs. In the context of currently available therapeutic options it is not considered ethical to withhold effective therapy from patients (i.e. receiving placebo only) for more than 3 months to study the efficacy of a new drug.[4] To address this issue it is common for an 'early escape' to be built into the study protocol. Patients randomized to the placebo arm are allowed to switch to active treatment in case of insufficient response after 12 or 16 weeks, or will leave the trial. In these patients, radiographs are taken at the moment of early escape. Due to the timescales required to typically observe structural changes, radiographic study endpoints are defined after (at least) 24 weeks follow-up. In patients who complete 24 week follow-up, the first follow-up radiographs are usually taken at this time point. Consequently, radiographs are available either at week 12/16 for a part of the patients (early escapers) or at week 24 radiographs for the patients completing according to protocol.

In order to adjust radiographic progression scores of early escapers (before week 24), linear extrapolation (LE) is a frequently used approach.[4] With LE, observed values are extrapolated beyond the last measurement,[5,6] under the assumption that radiographic change is a linear process on a group level.[7,8] An alternative approach is last observation carried forward (LOCF): the last observed value is used to impute subsequent missing data points.[5,6] Traditionally in trials LE is applied to radiographic data. LE is however sometimes criticized because it could lead to spurious overestimation of actual change in the placebo (or control) group that has most 'early escapers', and could therefore lead to overestimation of a treatment effect. LOCF, on the contrary, artificially reduces the change signal, and is for that reason sometimes preferred over LE as a conservative imputation method.

In the OSKIRA-1 trial,[9] used as an illustration here, radiographs were scheduled at week 12 in all patients still on study, regardless of early escape. Consequently, for patients completing 24 weeks, we have both week 12 and week 24 radiographic scores available. This provided an excellent opportunity to compare the performance of LOCF and LE with regard to radiographic change, in the context of the truly observed proportion of missingness, as well as in the context of modeled levels of missingness.

PATIENTS AND METHODS

Study design

Data from the phase 3 OSKIRA-1 trial were used.[9] This multicenter, randomized, double-blind, placebo-controlled parallel group study enrolled 918 patients with active rheumatoid arthritis despite current treatment with methotrexate to study the efficacy of fostamatinib. As no statistically significant effect on radiographic progression was demonstrated for fostamatinib

compared to placebo and as it was not the aim of this analysis to reassess the efficacy of fostamatinib, all arms were combined for the current analysis. The medical ethics committee of each center approved the study protocol and all patients gave written informed consent.

Study endpoints

Conventional radiographs of hands and feet were scheduled at baseline and in all patients still under follow-up at week 12, 24 and 52. The films were assessed independently by two trained readers using the Sharp/ van der Heijde score (SHS, range 0 – 448), blinded for patient identity, randomization arm and time order.[3] A third reader could adjudicate where required. The average score of the two readers was used. Radiographic change was defined as the difference in SHS between two subsequent time points.

Statistical analysis

Patients with complete radiographic data from baseline to week 24 were selected for this analysis. Observed change between the two time points was calculated. 1000 repeats of the dataset were taken with a random selection of 10% of the data set as missing at week 24 in each, creating 1000 replications of a treatment cohort with a known amount of missing data. This selection was made by the 'random sample of cases' function of SPSS. First, in each of the 1000 replications, missing data at week 24 were imputed with LE as follows: SHS at week 12 + change week 0 to 12, corrected for the actual days between radiographs. Second, missing data at week 24 in the 1000 variables were imputed using LOCF, i.e. using the week 12 result. This approach was repeated for the following percentages of missing data: 10, 20, 30, 40, 50, 60, 70, 80 and 90%. After imputation of the datasets with both LE and LOCF, for each created variable with missing and imputed data, the mean and standard deviations (SD) were calculated. Per proportion of missingness, the mean, minimum and maximum of the mean in 1000 estimates were calculated. Also, the mean, minimum and maximum of the SD in the 1000 estimates were calculated. Signal-to-noise ratios were provisionally estimated as mean of the means divided by the mean of the SD. Datasets obtained with percentages of LE and LOCF imputation were compared to the dataset with 100% truly observed data.

The total analysis was then repeated for patients with complete sets of radiographs from baseline to week 52. The estimates for LE were calculated as: SHS at week 24 + change week 0 to 24, corrected for the actual days between radiographs.

Patients were stratified according to disease duration at baseline, into 'early' disease (≤3 years; the lowest tertile) and 'established' disease (>3 years, the other tertiles). All analyses with 24 week follow-up data were repeated for both strata.

IBM SPSS statistics version 20 was used to perform the statistical analyses.

RESULTS

Of 918 patients randomized, 623 patients had complete radiographic data available from baseline to week 24. The majority of patients were female and middle-aged, with a median disease duration of 6 years (Table 1). Observed radiographic changes from for the several follow-up period and subgroups of patients are denoted in Table 2.

Table 1. Baseline characteristics of 623 patients included in the analysis for 24 week follow-up.

	Patients n=623
Female, n (%)	532 (85)
Age group, mean (SD)	52 (12)
RA duration in years, median (IQR)	6 (2 – 12)
DAS28-CRP, mean (SD)	5.8 (0.8)
SHS, mean (SD); median (IQR)	39 (52); 17 (4 – 54)
JSN, mean (SD); median (IQR)	21 (27); 9 (2 – 32)
Erosion score, mean (SD); median (IQR)	18 (27); 6 (2 – 22)

DAS28-CRP, disease activity score based on 28-joint count and C-reactive protein; IQR, interquartile range; JSN, joint space narrowing; RA, rheumatoid arthritis; SD, standard deviation; SHS, Sharp/ van der Heijde score (range 0 – 448).

Table 2. Observed change in Sharp/ van der Heijde score.

Follow-up period	**Patients**	**N**	**Mean**	**SD**	**Median**	**IQR**	**Min**	**Max**
Baseline – week 12	All patients	623	0.16	2.33	0.0	0.0 – 0.3	-11.0	49.0
Baseline – week 24	All patients	623	0.36	2.39	0.0	0.0 – 0.5	-5.0	49.0
Baseline – week 24	Disease duration ≤ 3 years	203	0.13	0.83	0.0	0.0 – 0.0	-3.0	5.5
Baseline – week 24	Disease duration > 3 years	420	0.46	2.85	0.0	0.0 – 0.5	-5.0	49.0
Baseline – week 52	All patients	534	0.74	2.98	0.0	0.0 – 0.5	-3.5	49.0

IQR, interquartile range; max, maximum; min, minimum; SD, standard deviation.

Table 3 shows the main results of the analysis with 24 weeks follow-up. For each percentage of missing data and for both imputation methods the mean and range (minimum to maximum) of the estimates of the mean across the 1000 replications is provided. The same is described for the standard deviations.

With a low percentage of missing data, both LE and LOCF methods estimate mean change scores close to the mean observed value of 0.36. However, by increasing percentages of missingness, LOCF results in an increasing underestimation of change (Figure 1A). Whilst the SD appears stable for the LOCF method in Figure 1B a small, but consistently ordered reduction in SD can be seen in Table 2 for the LOCF method, as might be expected if identical

Table 3. Change in Sharp/ van der Heijde score from baseline to week 24 in a simulation study with 1000 replications for each percentage of missing data imputed using linear extrapolation and last observation carried forward.

Mean (SD) observed SHS change was 0.36 (2.39).

Percentage missingness	Method	Mean of the estimated mean*	Mean of the estimated SD*	Lowest estimated mean*	Highest estimated mean*	Lowest estimated SD *	Highest estimated SD*	Signal-to-Noise ratio
	LE	0.36	2.65	0.28	0.48	2.36	4.34	0.14
10%	LOCF	0.34	2.39	0.28	0.38	2.30	2.45	0.14
	LE	0.35	2.88	0.21	0.49	2.39	4.40	0.12
20%	LOCF	0.32	2.38	0.26	0.36	2.30	2.46	0.13
	LE	0.35	3.17	0.21	0.49	2.43	4.49	0.11
30%	LOCF	0.30	2.37	0.22	0.36	2.30	2.45	0.13
	LE	0.34	3.36	0.21	0.50	2.43	4.53	0.10
40%	LOCF	0.28	2.37	0.20	0.34	2.27	2.45	0.12
	LE	0.34	3.57	0.19	0.48	2.53	4.55	0.09
50%	LOCF	0.26	2.36	0.19	0.32	2.28	2.44	0.11
	LE	0.33	3.83	0.17	0.47	2.61	4.59	0.09
60%	LOCF	0.24	2.35	0.16	0.31	2.27	2.43	0.10
	LE	0.33	4.02	0.19	0.50	2.64	4.62	0.08
70%	LOCF	0.22	2.35	0.16	0.30	2.27	2.43	0.09
	LE	0.33	4.26	0.19	0.44	2.78	4.63	0.08
80%	LOCF	0.20	2.34	0.16	0.26	2.26	2.43	0.09
	LE	0.32	4.45	0.20	0.40	2.90	4.64	0.07
90%	LOCF	0.18	2.34	0.14	0.23	2.26	2.41	0.08

*of 1000 replications with a randomly selected sample of missingness.
LE, linear extrapolation; LOCF, last observation carried forward; SHS, Sharp/ van der Heijde score; SD, standard deviation.

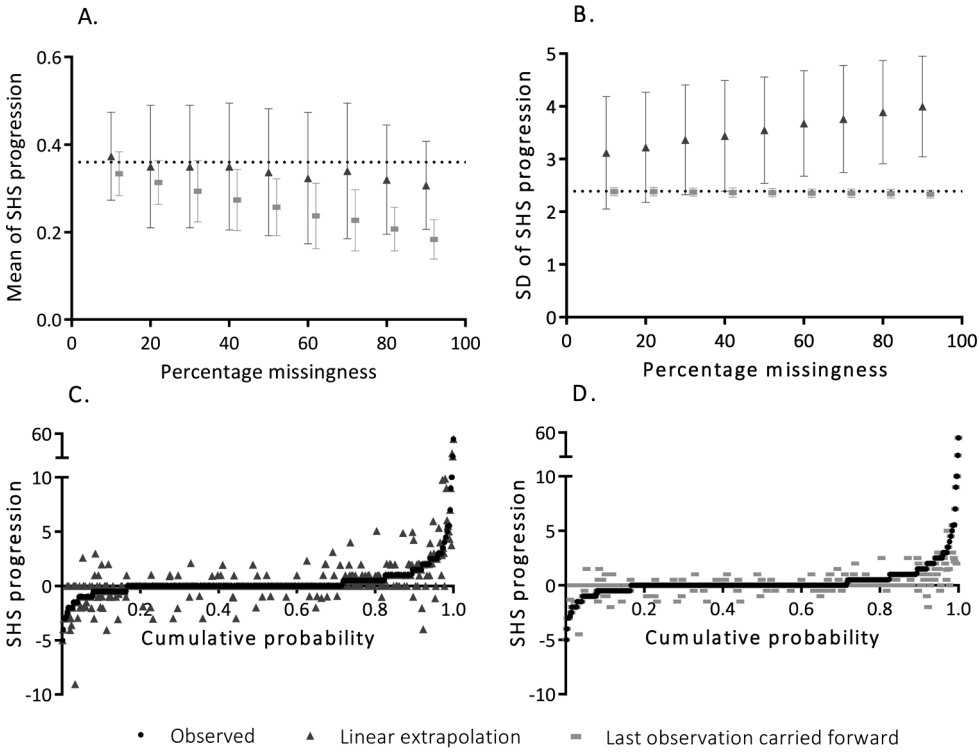


Figure 1: (A) The mean and range of all means of the 1000 replications, for each percentage of missingness, with 24 week follow-up. (B) The mean and range of all standard deviations of the replications, for each percentage of missingness, with 24 week follow-up. (C) Cumulative probability plot of radiographic change from baseline to week 24 with the complete observed data and linear extrapolation. (D) Cumulative probability plot of radiographic change from baseline to week 24 with the complete observed data and last observation carried forward.
Note Figure 1A and 1B: Dotted lines represent mean (1A) and standard deviation of observed change (1B).
Note Figure 1C and 1D: Cumulative probability plots represent paired samples of observed and imputed data, based on one replication (randomly chosen) with 50% percentage of missingness, in order to visualize the effect of the imputation methods. Statistics were based on 1000 replications and can therefore not be extracted from this plot.
SD, standard deviation; SHS, Sharp/van der Heijde score.

values are carried forward for more patients. In contrast, by increasing percentages of missing data, LE still gives a stable estimate of progression (with only small reductions at very large levels of missing data), but at the expense of increasing SD (Figure 1B). As a consequence, the 'signal-to-noise ratio' decreases for both methods with higher proportions missingness, and slightly more pronounced for LE, resulting in a higher signal-to-noise ratio for LOCF with more missing data. A cumulative probability plot, representing one sample with 50% missing data, is shown in Figure 1C and 1D to visualize the performance of LOCF and LE. These plots show that – in comparison with observed scores – LE leads to both some underestimation and overestimation in the tails of the distribution. However on the group level, the effect of this on the mean appears to even out. LOCF tends to reduce the amplitude of the distribution of

radiographic change (less extreme). Supplementary Figure 1 shows the same plot, but here both observed and imputed data are ranked independently, whereas Figure 1C and D show paired samples.

Results of the simulation study for the follow-up from baseline to week 52 are shown in Supplementary Table 1. For 52 week follow-up, a slight overestimation of the mean estimation by LE was seen. Again LOCF resulted in increasing underestimation by increasing percentages of missingness. If proportions of missingness increase, LE results in increasing SD, in contrast to a constantly low (and marginally decreasing) SD for LOCF.

For patients with a disease duration ≤3 years, LE results in a subtle underestimation, while LOCF results in an obvious underestimation of true observed change (Supplementary Table 2). Once more, the SD of LOCF only shows very minor reductions with increasing percentage missingness, whereas the SD of LE increases with more missing data. For patients with >3 years disease duration, estimates using LE are close to the observed value with inflating SD as missingness increases (Supplementary Table 3). With LOCF, mean SHS change is estimated lower than observed, an effect becoming more explicit with high proportions of missingness, with a stable (slightly reducing) SD.

DISCUSSION

In this study, we have compared two 'custom' methods for imputing missing data on radiographic change for clinical trials with rheumatoid arthritis. We have demonstrated that linear extrapolation generally results in accurate and stable estimates of mean change at the group level, even with increasing proportions of missingness. However, with a high proportion of missingness, the SD of the estimates substantially inflates. It is natural that variation and uncertainty should increase alongside the amount of missing data. As expected, last observation carried forward results in progressive underestimation of change, with a marginal reduction in SD with increasing percentages missingness. This modeling experiment suggests that LE generally does not lead to overestimation of radiographic change.

Joint damage progression assessed on conventional radiographs of hands and feet comprises an important benchmark in the treatment of rheumatoid arthritis. Handling missing radiographic data has always been a challenge. Although both LE and LOCF are widely used, validating studies encompassing both methods have never been performed.

With a complete radiographic dataset of 623 patients for 24 week follow-up, and of 534 patients for 1 year follow-up, we had the opportunity to perform this validation study. By intentionally making a percentage of radiographic data missing, we were able to compare both imputation methods, using known levels of missingness, to the true, observed data.

LOCF resulted in progressive underestimation of actual change when proportions of missingness increased. The SD remained relatively stable with only small decreases with the percentage of missingness, which can be explained by the fact that more zeros were imputed using LOCF. LE resulted in accurate and stable mean estimates that remained close

to the population mean regardless of the proportion of missing data. This is an important observation as in a placebo controlled study with early escape it is to be expected that there would be a greater proportion of missing data in a placebo arm than an experimental treatment arm. As such any method that is affected by the proportion of missing data would introduce bias when comparing treatment groups in a randomized controlled study. Nevertheless, increasing missingness is accompanied by inflating SD when using LE and the reducing signal-to-noise ratio indicates that identifying a treatment effect may become more challenging as the proportion of missing data increases. In fact the Type I error (inappropriately rejecting the null hypothesis of no difference between both trial arms) may be reduced for both methods, but for differing reasons. Although LOCF has been considered as 'the more conservative approach' in the past this simulation study shows that this is due to a consistent underestimation of the level of progression and a slight artificial reduction of the standard deviation. This underestimation of progression for a placebo arm with a high degree of missing data (sometimes up to 60%) would lead to Type II errors (inappropriately accepting the null hypothesis of no difference). Whilst the power to demonstrate treatment effects would also be reduced using the LE method this simulation study suggests that estimates of the mean change in SHS are likely to remain more robust. This finding was consistent for 24 week and 52 week follow-up, for patients with a short disease duration (≤3 years) and a longer disease duration (>3 years). This makes LE an appropriate method to apply to missing radiographic change data after an early escape.

A weakness of our study is that the MCAR scenario of 'missing completely at random' is likely not representative for what is happening in a clinical trial, in which differential drop out may occur and in which patients with worse prognosis (higher disease activity) may more often use the early escape option. The extrapolation methods used here do account for observed values and so in a sense go someway to making a 'missing at random' (MAR) assumption. While we have not modeled scenarios with 'missing not at random' (so called MNAR scenarios) we have captured in this study many patients with relatively high progression scores over 24 weeks, ensuring that this type of patient is included in the simulation project. In fact, such an MNAR scenario could likely imply that early escapers (preferentially occurring in the placebo arm) would have a progression that in truth is worse than the progression signal assumed by LE: in such patients LE is a conservative estimate of true progression. We therefore do not think that MNAR scenarios will provide results that are in contrast with our conclusions.

The performance of both imputation methods under study were tested with up to 90% missing data. Even though LE gives an accurate impression of change estimates, percentages missingness of this magnitude are obviously undesirable and never encountered in reality, and we have performed this only to test extreme situations.

In conclusion, linear extrapolation performs well in estimating radiographic change, irrespective of the percentage of missingness, and does not inappropriately increase the likelihood of Type I error. In contrast, last observation carried forward results in a consistent and progressive underestimation of progression with spuriously low variability. In contrast

to what many think, linear extrapolation does not appear to spuriously overestimate radiographic progression and should therefore be preferred over LOCF as it provides a more accurate estimate of the mean change in SHS in the presence of missing data.

REFERENCES

1 Sharp JT, Lidsky MD, Collins LC, et al. Methods of scoring the progression of radiologic changes in rheumatoid arthritis. Correlation of radiologic, clinical and laboratory abnormalities. ***Arthritis Rheum*** 1971;14:706–20.

2 Sharp JT, Young DY, Bluhm GB, et al. How many joints in the hands and wrists should be included in a score of radiologic abnormalities used to assess rheumatoid arthritis? ***Arthritis Rheum*** 1985;28:1326–35.

3 Van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. ***J Rheumatol*** 1999;26:743–5.

4 Landewé R, Strand V, van der Heijde D. From inhibition of radiographic progression to maintaining structural integrity: a methodological framework for radiographic progression in rheumatoid arthritis and psoriatic arthritis clinical trials. ***Ann Rheum Dis*** 2013;72:1113–7.

5 Van der Heijde D, Landewé R, Klareskog L, et al. Presentation and analysis of data on radiographic outcome in clinical trials: Experience from the TEMPO study. ***Arthritis Rheum*** 2005;52:49–60.

6 Landewé R, van der Heijde D. Presentation and analysis of radiographic data in clinical trials and observational studies. ***Ann Rheum Dis*** 2005;64:iv48–51.

7 Hulsmans HMJ, Jacobs JWG, van der Heijde DMFM, et al. The course of radiologic damage during the first six years of rheumatoid arthritis. ***Arthritis Rheum*** 2000;43:1927–40.

8 Wolfe F, Sharp JT. Radiographic outcome of recent-onset rheumatoid arthritis: a 19-year study of radiographic progression. ***Arthritis Rheum*** 1998;41:1571–82.

9 Weinblatt ME, Genovese MC, Ho M, et al. Effects of Fostamatinib, an Oral Spleen Tyrosine Kinase Inhibitor, in Rheumatoid Arthritis Patients With an Inadequate Response to Methotrexate: Results From a Phase III, Multicenter, Randomized, Double-Blind, Placebo-Controlled, Parallel-Group Study. ***Arthritis Rheumatol*** 2014;66:3255–64.

Supplementary Table 1. Change in Sharp/ van der Heijde score from baseline to week 52 in a simulation study with 1000 replications for each percentage of missing data imputed using linear extrapolation and last observation carried forward.

Mean (SD) observed SHS change was 0.74 (2.98).

Percentage missingness	Method	Mean of the estimated mean*	Mean of the estimated SD*	Lowest estimated mean*	Highest estimated mean*	Lowest estimated SD *	Highest estimated SD*	Signal-to-Noise ratio
	LE	0.76	3.23	0.66	0.92	2.84	5.16	0.23
10%	LOCF	0.71	2.94	0.61	0.76	2.73	3.00	0.24
	LE	0.77	3.52	0.63	0.95	2.88	5.24	0.22
20%	LOCF	0.68	2.90	0.59	0.75	2.68	2.99	0.23
	LE	0.78	3.73	0.62	0.97	2.90	5.28	0.21
30%	LOCF	0.64	2.85	0.54	0.75	2.60	2.99	0.23
	LE	0.79	3.99	0.62	0.96	2.94	5.37	0.20
40%	LOCF	0.61	2.79	0.51	0.71	2.59	2.98	0.22
	LE	0.80	4.25	0.62	0.99	2.95	5.38	0.19
50%	LOCF	0.57	2.75	0.46	0.71	2.53	2.95	0.21
	LE	0.81	4.43	0.62	0.98	3.00	5.39	0.18
60%	LOCF	0.54	2.70	0.44	0.63	2.49	2.92	0.20
	LE	0.82	4.71	0.64	1.01	3.07	5.42	0.17
70%	LOCF	0.50	2.65	0.41	0.61	2.48	2.90	0.19
	LE	0.83	4.91	0.65	0.96	3.21	5.40	0.17
80%	LOCF	0.47	2.59	0.39	0.55	2.47	2.82	0.18
	LE	0.84	5.15	0.67	0.95	3.27	5.42	0.16
90%	LOCF	0.43	2.54	0.38	0.51	2.47	2.79	0.17

*of 1000 replications with a randomly selected sample of missingness

LE, linear extrapolation; LOCF, last observation carried forward; SHS, Sharp/ van der Heijde score; SD, standard deviation.

Supplementary Table 2. Change in Sharp/ van der Heijde score from baseline to week 24 in patients with a disease duration ≤3 years, in a simulation study with 1000 replications for each percentage of missing data imputed using linear extrapolation and last observation carried forward.

Mean (SD) observed SHS change was 0.13 (0.83).

Percentage missingness	Method	Mean of the estimated mean*	Mean of the estimated SD*	Lowest estimated mean*	Highest estimated mean*	Lowest estimated SD *	Highest estimated SD*	Signal-to-Noise ratio
	LE	0 .13	0 .93	0 .03	0 .21	0 .80	1 .22	0 .14
10%	LOCF	0 .13	0 .83	0 .06	0 .17	0 .71	0 .91	0 .15
	LE	0 .13	1 .02	-0 .01	0 .24	0 .80	1 .38	0 .13
20%	LOCF	0 .12	0 .82	0 .05	0 .18	0 .68	0 .93	0 .14
	LE	0 .12	1 .11	0 .00	0 .23	0 .82	1 .45	0 .11
30%	LOCF	0 .11	0 .82	0 .03	0 .18	0 .62	0 .94	0 .13
	LE	0 .12	1 .19	-0 .01	0 .24	0 .90	1 .50	0 .10
40%	LOCF	0 .10	0 .82	0 .03	0 .17	0 .66	0 .94	0 .12
	LE	0 .12	1 .26	-0 .02	0 .23	0 .98	1 .53	0 .09
50%	LOCF	0 .09	0 .81	0 .02	0 .17	0 .63	0 .95	0 .11
	LE	0 .11	1 .33	0 .00	0 .24	0 .97	1 .54	0 .08
60%	LOCF	0 .08	0 .81	0 .02	0 .15	0 .66	0 .94	0 .10
	LE	0 .11	1 .40	-0 .01	0 .24	1 .09	1 .59	0 .08
70%	LOCF	0 .07	0 .81	0 .01	0 .15	0 .67	0 .94	0 .09
	LE	0 .10	1 .47	0 .00	0 .20	1 .15	1 .60	0 .07
80%	LOCF	0 .06	0 .81	0 .01	0 .13	0 .68	0 .93	0 .08
	LE	0 .10	1 .53	0 .02	0 .19	1 .28	1 .60	0 .06
90%	LOCF	0 .05	0 .80	0 .02	0 .10	0 .71	0 .90	0 .07

*of 1000 replications with a randomly selected sample of missingness

LE, linear extrapolation; LOCF, last observation carried forward; SHS, Sharp/ van der Heijde score; SD, standard deviation.

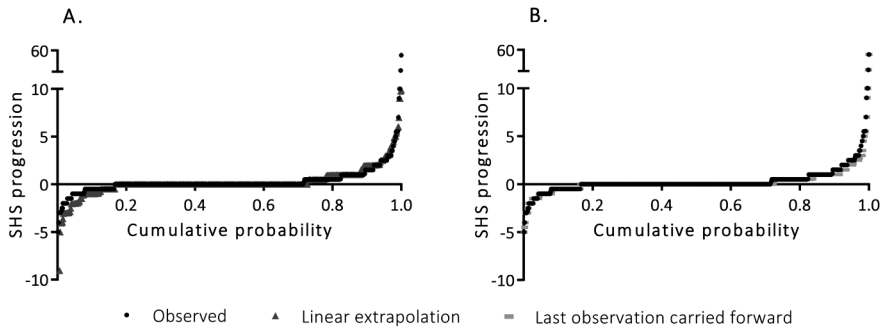
Supplementary Table 3. Change in Sharp/ van der Heijde score from baseline to week 24 in patients with a disease duration >3 years, in a simulation study with 1000 replications for each percentage of missing data imputed using linear extrapolation and last observation carried forward.

Mean (SD) observed SHS change was 0.46 (2.85).

Percentage missingness	Method	Mean of the estimated mean*	Mean of the estimated SD*	Lowest estimated mean*	Highest estimated mean*	Lowest estimated SD *	Highest estimated SD*	Signal-to-Noise ratio
	LE	0 .46	3 .18	0 .32	0 .63	2 .80	5 .28	0 .15
10%	LOCF	0 .44	2 .84	0 .37	0 .50	2 .75	2 .93	0 .16
	LE	0 .46	3 .47	0 .26	0 .65	2 .81	5 .34	0 .13
20%	LOCF	0 .42	2 .83	0 .32	0 .49	2 .73	2 .93	0 .15
	LE	0 .45	3 .72	0 .27	0 .65	2 .86	5 .38	0 .12
30%	LOCF	0 .39	2 .83	0 .30	0 .46	2 .71	2 .93	0 .14
	LE	0 .45	4 .06	0 .24	0 .70	2 .88	5 .43	0 .11
40%	LOCF	0 .37	2 .83	0 .27	0 .45	2 .71	2 .93	0 .13
	LE	0 .45	4 .32	0 .24	0 .64	2 .92	5 .46	0 .10
50%	LOCF	0 .34	2 .82	0 .24	0 .44	2 .70	2 .93	0 .12
	LE	0 .44	4 .52	0 .24	0 .62	2 .96	5 .50	0 .10
60%	LOCF	0 .32	2 .81	0 .22	0 .42	2 .71	2 .93	0 .11
	LE	0 .44	4 .81	0 .24	0 .65	3 .11	5 .53	0 .09
70%	LOCF	0 .29	2 .81	0 .22	0 .38	2 .70	2 .91	0 .10
	LE	0 .43	5 .06	0 .21	0 .60	3 .17	5 .55	0 .09
80%	LOCF	0 .27	2 .80	0 .21	0 .35	2 .69	2 .91	0 .10
	LE	0 .43	5 .30	0 .26	0 .57	3 .38	5 .55	0 .08
90%	LOCF	0 .24	2 .79	0 .20	0 .31	2 .70	2 .87	0 .09

*of 1000 replications with a randomly selected sample of missingness

LE, linear extrapolation; LOCF, last observation carried forward; SHS, Sharp/ van der Heijde score; SD, standard deviation.



Supplementary Figure 1. Cumulative probability plot of radiographic change from baseline to week 24 with the complete observed data and imputed data, both independently ranked from the lowest to the highest value. (A) Complete observed data and linear extrapolation. (B) Complete observed data and last observation carried forward. Note: Cumulative probability plots represent independently ranked samples of observed and imputed data, based on one replication (randomly chosen) with 50% percentage of missingness, in order to visualize the effect of the imputation methods. Figure 1 depicts the same replication variable with paired samples. Statistics were based on 1000 replications and can therefore not be extracted from this plot.
SD, standard deviation; SHS, Sharp/van der Heijde score.