Title: Imputation techniques for incomplete radiographic outcome in rheumatoid arthritis randomized trials: a simulation study

Authors: Gabriel Baron[1], Adeline Samson[1], Bruno Giraudeau[2],Isabelle Boutron[1], Maxime Dougados[3], Philippe Ravaud[1]

[1]Gabriel Baron, MSc, Isabelle Boutron, MD, Philippe Ravaud, MD, PhD: INSERM U738, Paris, France ; Université Paris 7 Denis Diderot, UFR de Médecine, Paris, France ; AP-HP, Hôpital Bichat, Département d'Epidémiologie, Biostatistique et Recherche Clinique, Paris , France ;

[2]Bruno Giraudeau, PhD: INSERM CIC 202, Tours, France.

[3]Paris-Descartes University, Medecine Faculty, Paris , France, AP-HP, Cochin Hospital, 27 rue du Faubourg Saint Jacques, 75014 Paris, France

Address correspondence and reprint requests to

Gabriel Baron, Département d'Epidémiologie Biostatistique et Recherche Clinique, INSERM U738, Groupe Hospitalier Bichat-Claude Bernard, 46 rue Henri Huchard, 75018 Paris, France.

Tel: +33 01 40 25 62 57

Fax: +33 01 40 25 67 73

E-mail: gabriel.baron@bch.aphp.fr

Word count:

ABSTRACT

Objective.

Methods.

Results.

Conclusion.

Key words: rheumatoid arthritis, randomized controlled trials, simulation study, radiographic outcome, missing data

Rheumatoid arthritis (RA) is the most common chronic inflammatory joint disease and is responsible for symptomatic manifestations (e.g., functional status, pain) and structural damage (i.e., damage of the articular cartilage and bone) (1). The use of disease-modifying anti-rheumatic drugs has increased for RA (2). Assessing such treatments requires the measurement of structural outcomes in randomized controlled trials to demonstrate a reduction or a retardation of disease progression. Radiography provides an objective measure of the extent of anatomical joint damage. It can be used to assess the severity of the structural destruction, to follow the course of the disease and to establish effects of treatment (3). The assessment of radiographic outcomes for evaluating drug efficacy was recommended for the management of RA in controlled trials (4, 5) and the radiographic outcome is often used as a primary endpoint for assessing structural severity (6).

The intention-to-treat (ITT) principle is the cornerstone of superiority trials (7-9) and is widely recommended (10, 11). The ITT principle requires that all patients, whether complete or incomplete, be included in the statistical analysis. The application of this principle in trials evaluating radiographic outcomes in RA is low (12). In these trials involving longitudinal measurements of radiographic outcome, missing data can appeared from reasons such as lack of efficacy or adverse events. When data are incomplete, results of the trial can be affected in two major ways. First, a bias of treatment effect estimates due to missing data may appeared. For example, patients who are experiencing more deterioration may be less likely to complete the visits. If missing data are ignored and analyses are based only on the data of patients who are doing well, then the disease progression could be underestimated (13). The second consequence is the loss power because of reduced sample size if some subjects are completely excluded from the analysis.

Conclusions of trials may be affected by the method of handling used. Here is the reason why it is largely advised to perform sensitivity analyses (i.e., are way of handling missing data influence conclusions of the study?) in order to be sure that the qualitative conclusion of a randomized trial is not affected by the way missing data are handled. Recently, 2 sensitivity analyses, which evaluated different methods of handling radiographic missing data, were performed to confirm the robustness of radiographic results of published trials in RA (14, 15). However, although sensitivity studies should be entire part of the statistical analysis plan of a randomized study, they do not allow to draw general conclusions (i.e., conclusions applicable to different trials) regarding the most appropriate method to be used to deal with missing data for the main principal analysis. This is precisely the aim of this

simulation study. We proposed to compare some data analysis strategies and imputation techniques on simulated trials with radiographic outcome.

METHODS

The goal of this study was to evaluate the validity of several commonly used approaches to dealing with missing data under a scenario that mimic actual RA trials with radiographic outcome. We simulated trials on which "ideal" analysis (i.e., complete information for all subjects) could be performed and trials with missing data according to a missingness mechanism. Considered data analysis were T Test and linear mixed-effect models. Investigating the treatment of missing data involved the case-complete approach, last observation carried forward (LOCF) approach and multiple imputation. Results will focus on type I error, power of different approaches and on the magnitude of bias introduced by missing data.

The underlying clinical trial

A simulation based on 2-armed randomized controlled trials resembling that in RA trials with was considered (a control group and an experimental group). A two-year duration trial with 2 time points of measurement starting from baseline was assumed (all time increments were equal to 1 year).

The primary endpoint was the Sharp-Van der Heijde score (16, 17), a quantitative radiological measure which is actually recommended to be one of the 2 possible primary endpoint when evaluating structural damage (18). This score assesses erosions and joint-space narrowing separately in the hands and feet and ranges from 0 to 448. Thirty-two joints in the hands and 12 in the feet are scored for erosions, with a maximum of 5 erosions per joint in the hands and 10 in the feet. Joint-space narrowing is graded from 0 to 4 in 30 joints in the hands and in 12 joints in the feet. The Sharp-Van der Heijde score is the sum of the erosion score and the joint-space narrowing score.

Simulations of longitudinal measurements were performed using a linear mixed effect model with random intercept and slope. This model fits an intercept and slope for each patient's damage score over time. According to published data of the TEMPO study (14, 19, 20), we assumed that the baseline distribution of the radiological scores can be approximated by a log normal approximation (mean=45, standard deviation=45). The mean progression can be assumed to be linear (21) although individual patient's evolution show high variability (22). The slope (and its standard deviation) was simulated by a normal distribution. Under the alternative hypothesis, the slope and its standard deviation were assumed to be more

important in the control group than in the experimental group (mean change over 2 years=3, standard deviation=10 versus mean=0, standard deviation 5), reflecting lesser benefits from treatment and so higher deterioration of structural damage. This scenario is approximate (expected mean change between the 2 groups=3, common standard deviation=7.9, effect size=0.38) but nevertheless reflects a tendency in RA trials. This simulation study also investigates possible consequences on type I error. Under the null hypothesis, mean change over 2 years was equal to 0 in each group.

Missingness mechanism

After the complete data sets are created, patient data were deleted on the basis of the following considerations. We only consider missing data with a monotone pattern (i.e., data for a patient up to a certain time). We assume that all the baseline data are observed. Further to our previous literature review, we assume a plausible rate of missing data equal to 20% in both group (experimental and control) at 2 years. Patients with disease progression greater or lower than a defined limit between 2 occasions dropped out the trial with a probability equal to 2.5% if the slope between 2 successive visits was negative (i.e., improvement), 5% if the slope was between 0 and 5 points (i.e., slight deterioration) and 20% if the slope was greater then 5 points (i.e., important deterioration). The limit of 5 points was chosen in accordance with published estimation of minimally clinically important difference and smallest detectable difference of the Sharp-Van der Heijde score which are very close (around 5 points) (17). Probability of missing value were arbitrarily chosen to ensure global dropout rate of around 20% at the final visit. In a second scenario, probability of dropout were divided by 2 in order to have a dropout rate around 10%.

Methods of management of missing data

We consider 3 methods of management. The first method, which ignores the problem of missing data, is the case-complete analysis, which uses only patients with complete data. The second method was the LOCF method, the most popular method of single imputation. According to this method, the last observation is carried forward and is used for all missing observations at the remaining time points. This method was not applied when only the baseline visit was available. The third method was multiple imputation. Instead of filling in a single value for each missing value, this technique replaces each missing value of an

incomplete dataset by a set of plausible value that represent the uncertainty about the right value to impute. Each completed dataset is analyzed by the analysis of choice and results of imputed datasets are combined in a single analysis yielding point estimates and standard errors.

Data analysis

The 2-sided T-test was used to test the absolute change between the 2 groups (evolution estimated by the change between the baseline visit and the visit 2 years after). A linear model with mixed effects for repeated measurements with random intercept and slope was also considered. This method exploits the richness of the dynamic obtained by repeated measurements by performing restricted maximum likelihood estimation with all available data. Time of the visit and group by visit interaction were fixed effect and F-tests of fixed effect were computed. These data analysis strategies were applied alone (without data handling) but also after application of LOCF and multiple imputation (table 1). P values less than 0.05 were considered significant. All these approaches will be applied to scenarios of missing data.

Results of simulation

The power of applying different approaches for dealing with missing values was computed. To estimate the empirical power of each approach, the entire trial simulation was repeated 1000 times under the alternative hypothesis. The percentage of these 1000 separate simulated trials in which test was statistically significant (at the 5% level) was the estimated power of the trial for that approach. The empirical type I error was calculated as the proportion of p-values from testing the null hypothesis of no difference on each simulated trials that are less than the nominal 5% level, when the null hypothesis is true.

The results denoted by $\hat{\vartheta}_i$ (i.e., estimators of treatment effect and its standard deviation computed on each simulated trials *i*) and obtained by each approach under consideration on simulated data sets (complete and incomplete) were assessed by their comparison with the expected results (i.e., true parameters denoted by θ^o). For comparisons, we have considered

the relative bias (RB= $100 * \frac{\frac{1}{n}\sum_i (\hat{\theta}_i - \theta^o)}{\theta^o}$) and the relative root mean square error (RMSE= $100 * \frac{\sqrt{\frac{1}{n}\sum_i (\hat{\theta}_i - \theta^o)^2}}{\vartheta^o}$) of the treatment effect and its standard deviation for each approach involved in the comparison (where *n* is the number of simulated trials, i.e., 1000).

Simulations were conducted using R and statistical analyses were performed by using SAS version 9.1 software.

RESULTS

Simulated data look like published figures (figure 1). The missing data rate at 2 years were equal to 22% in the first simulated scenario and to 11% in the second.

As expected, RB of treatment effect on complete data sets were low whatever the method of data analysis (table 2). RB of standard error estimated by linear mixed effect model were very low.

When considering a rate of missing data around 22%, results in table 2 show the following (table 2): 1) The linear mixed effect model rendered the most precise results for both treatment effect and standard deviation. For instance, RB of treatment effect of estimated by linear mixed effect model was equal or lesser then RB with a T Test approach whatever the methods of data handling.

2) The multiple imputation was much precise than LOCF approach but less than case complete approach.

3) The linear mixed effect model applied on all available data was the most precise method.

Under the null hypothesis, type I error were preserved allowing us to compare power under the alternative hypothesis. When using LOCF, power was higher than for other imputation strategies. When considering the linear mixed effect model, LOCF method and analysis of all available cases are equivalent. However, the loss of power was around 20% in comparison with analysis on complete data set.

Results provided by a 11% rate of dropout are less spectacular but follows the same conclusions (table 2).

DISCUSSION

This work was concerned with comparison of imputation techniques applied to incomplete longitudinal data sets in the field of radiographic measurements in RA. The data sets were simulated to resemble time behaviour of radiographic outcome in RA randomized controlled trials. The missingness mechanisms employed resembled the process of withdrawal from trials due to lack of efficacy. The analysis compared included T Test and linear mixed effect model. Imputation techniques compared included the case complete approach, the LOCF method and multiple imputation. The results of simulation shows that the linear mixed effect model applied on case complete patients give in average, a better power and more precise estimations of means and standard deviation for treatment effect than the other methods under comparisons. However when rate of missing data is around 20%, the loss of power is important (around 20%) and bias on treatment effect can not be neglected (around 13%).

The problem of dealing with missing data is tackled extensively in methodological works involving radiographic endpoints in rheumatoid arthritis (23-25). To our knowledge, a simulation approach was already used in osteoporosis (26), an another progressively deteriorating disease, to investigate consequences on type 1 error and power of applying different methods for dealing with missing data, but not yet in rheumatoid arthritis .

It should be stressed that methods minimizing missing data by a well designed study is the first issue to consider this problem (e.g., appropriate follow-up for all randomized patients: schedule for a radiographic visit even if the patient drop out the study) (27). Statistical methods, however, well designed, cannot address missing values when their proportion is particularly high. In this work, the proportion of missing values was sufficiently low (i.e., 10 or 20%) so as to be considered reasonably with statistical methods.

In rheumatoid arthritis trials involving longitudinal measurements of radiograhic outcomes, 2 main sources of missing data are identified: lack of efficacy and adverse events. Unexpected selective dropout (preferentially in one group) due to lack of clinical efficacy may bias the trial results. In general, patients with a worse prognosis (greater disease activity, greater evidence of progression seen on radiology) have a higher prior probability of premature discontinuation in any clinical trial, and patients completing the entire trial have a more favourable prognosis, either by nature or by treatment (29). When dropout results from disease progression or therapeutic ineffectiveness, missing data could depends on earlier observations (i.e., missing at random) or on radiological score at the time the assessment is

missing (i.e., missing at random) (30). Missing at random data are plausible in true well-controlled studies, such as clinical trials in which extensive efforts are made to observe all the outcomes and the factors that influence them (28). More generally, the missing at random assumption in a longitudinal study is more likely to be valid if missingness can be well explained by observed data (e.g., previous radiological scores or baseline prognostic factors) (31). Hence, clinical trials by their very design seek to minimize the amount of missing not at random data (28). In our study, we particularly focused on missing data due to lack of efficacy by excluding patients having a deterioration with a high probability of dropout. Adverse events can lead patients to leave the trial independently of good or bad treatment results. Consequently, we can assume that missing data due to adverse events is related to the group of treatment (28).

Calculating mean change in each group between baseline and the end of the study and comparing it by the classical t test was the most popular method of analysis. However, because of trials designed with longitudinal measurements, the use of appropriate statistical methods for repeated measurements is now recommended (14). We chosen the linear mixed effect model which tends to be more powerful than classical T test. This model takes into account how the disease and treatment affects each patient over time and how the radiographic data of the same patient are correlated.

Case complete approach is conflicting with the intention to treat principle and is now more and more avoided. However, use of this approach allowed us to have a reference to quantify bias introduced by missing data which can not be neglected.

With the LOCF approach, the missing radiographic value is replaced by the last available value, assuming no change in radiographic score after the dropout. In rheumatoid arthritis trials, this concern lead to considering dropouts as having smaller radiographic damage than completers. This approach is widely criticized and unsurprisingly introduced bias in the estimates of longitudinal changes and underestimates standard deviation.

Contrary to LOCF approach, multiple imputation have theoretically good statistical properties (e.g. unbiased estimates) provided that data are missing at random. In this study, multiple imputation did not bring substantially improvement on power and estimation of treatment effect when compared with case complete approach. However, we can easily imagine that the multiple imputation can give better results with a dataset containing demographic data and predictors of the severity of the disease for instance.

Using linear mixed effect models do not handle missing values, but estimates take into account all available data whether all longitudinal measurements are complete or not. If the

data are missing at random then the estimates will be theoretically unbiased. In our study, this method applied without imputation strategy minimized both problem related to power and treatment effect. In our study we found that this method is particularly interesting when the number of dropout become large (around 20%).

These study also has caveats and limitations. The random simulations carried out may not be reflective of the patterns of missing data seen in real situations. Furthermore, we do not explore more sophisticated methods of dealing with missing data such as selection models or pattern mixture models. However, because these models relies on many assumptions, sensitivity analyses are advised and so they can not be used as a main strategy when analyzing a randomized controlled trial.

In this study we showed the influence of the choice of strategies of analysis and methods for handling missing data when designing clinical trials with radiographic outcomes. Our results, especially those obtained with the use of linear mixed effect models can help investigators in planning clinical trials, especially when choosing methods of data analysis and when designing sensitivity analysis. Unfortunately, this method partially correct power and bias and so efforts to minimize missing data should be encouraged.

References

1. Symmons DP. Disease assessment indices: activity, damage and severity. Baillieres Clin Rheumatol 1995;9:267-85.
2. Lee DM, Weinblatt ME. Rheumatoid arthritis. Lancet 2001;358:903-11.
3. Soubrier M, Dougados M. How to assess early rheumatoid arthritis in daily clinical practice. Best Pract Res Clin Rheumatol 2005;19:73-89.
4. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.
5. Recommendations for the registration of drugs used in the treatment of rheumatoid arthritis. Group for the Respect of Ethics and Excellence in Science (GREES): rheumatoid arthritis section. Br J Rheumatol 1998;37:211-5.
6. Boers M, van der Heijde DM. Prevention or retardation of joint damage in rheumatoid arthritis: issues of definition, evaluation and interpretation of plain radiographs. Drugs 2002;62:1717-24.
7. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. Stat Med 1999;18:1903-42.
8. Lewis JA, Machin D. Intention to treat--who should use ITT? Br J Cancer 1993;68:647-50.
9. Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000;21:167-89.
10. CPMP CfPMP. ICH Topic E9 Statistical principles for clinical trials. The European Agency for the Evaluation of Medicinal Products 1998;CPMP/ICH/363/96.
11. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663-94.
12. Baron G, Boutron I, Giraudeau B, Ravaud P. Violation of the intent-to-treat principle and rate of missing data in superiority trials assessing structural outcomes in rheumatic diseases. Arthritis Rheum 2005;52:1858-1865.
13. CPMP CfPMP. Points to consider on missing data. The European Agency for the Evaluation of Medicinal Products 2001;CPMP/EWP/1776/99.

14. van der Heijde D, Landewe R, Klareskog L, Rodriguez-Valverde V, Settas L, Pedersen R, et al. Presentation and analysis of data on radiographic outcome in clinical trials: experience from the TEMPO study. Arthritis Rheum 2005;52:49-60.

15. Leung H, Hurley F, Strand V. Issues involved in a metaanalysis of rheumatoid arthritis radiographic progression. Analysis issues. J Rheumatol 2000;27:544-8; discussion 552.

16. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27:261-3.

17. Bruynesteyn K, van der Heijde D, Boers M, van der Linden S, Lassere M, van der Vleuten C. The Sharp/van der Heijde method out-performed the Larsen/Scott method on the individual patient level in assessing radiographs in early rheumatoid arthritis. J Clin Epidemiol 2004;57:502-12.

18. van der Heijde D, Simon L, Smolen J, Strand V, Sharp J, Boers M, et al. How to report radiographic data in randomized clinical trials in rheumatoid arthritis: guidelines from a roundtable discussion. Arthritis Rheum 2002;47:215-8.

19. Klareskog L, van der Heijde D, de Jager JP, Gough A, Kalden J, Malaise M, et al. Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. Lancet 2004;363:675-81.

20. van der Heijde D, Klareskog L, Rodriguez-Valverde V, Codreanu C, Bolosiu H, Melo-Gomes J, et al. Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the TEMPO study, a double-blind, randomized trial. Arthritis Rheum 2006;54:1063-74.

21. Hulsmans HM, Jacobs JW, van der Heijde DM, van Albada-Kuipers GA, Schenk Y, Bijlsma JW. The course of radiologic damage during the first six years of rheumatoid arthritis. Arthritis Rheum 2000;43:1927-40.

22. Plant MJ, Jones PW, Saklatvala J, Ollier WE, Dawes PT. Patterns of radiological progression in early rheumatoid arthritis: results of an 8 year prospective study. J Rheumatol 1998;25:417-26.

23. Symmons DP. Methodological issues in conducting and analyzing longitudinal observational studies in rheumatoid arthritis. J Rheumatol Suppl 2004;69:30-4.

24. Johnson K. Evidence from rheumatoid arthritis trials for approval: what does it mean? With special reference to using radiographic endpoints. J Rheumatol 2000;27:549-51.

25. Landewe R, van der Heijde D. Presentation and analysis of radiographic data in clinical trials and observational studies. Ann Rheum Dis 2005;64 Suppl 4:iv48-51.

26. Unnebrink K, Windeler J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Stat Med 2001;20:3931-46.
27. Auleley GR, Giraudeau B, Baron G, Maillefert JF, Dougados M, Ravaud P. The methods for handling missing data in clinical trials influence sample size requirements. J Clin Epidemiol 2004;57:447-53.
28. Mallinckrodt CH, Sanger TM, Dube S, DeBrota DJ, Molenberghs G, Carroll RJ, et al. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. Biol Psychiatry 2003;53:754-60.
29. Landewe RB, Boers M, van der Heijde DM. How to interpret radiological progression in randomized clinical trials? Rheumatology (Oxford) 2003;42:2-5.
30. Fairclough D. Design and analysis of quality of life studies in clinical trials: Chapman & Hall/CRC; 2002.
31. Neuenschwander B, Branson M. Modeling missingness for time-to-event data: a case study in osteoporosis. J Biopharm Stat 2004;14:1005-19.

Figure 1. Cumulative distribution of total Sharp Van der Heijde score over 2 years of treatment in each group (experimental and control) in one of the 1000 simulated datasets.

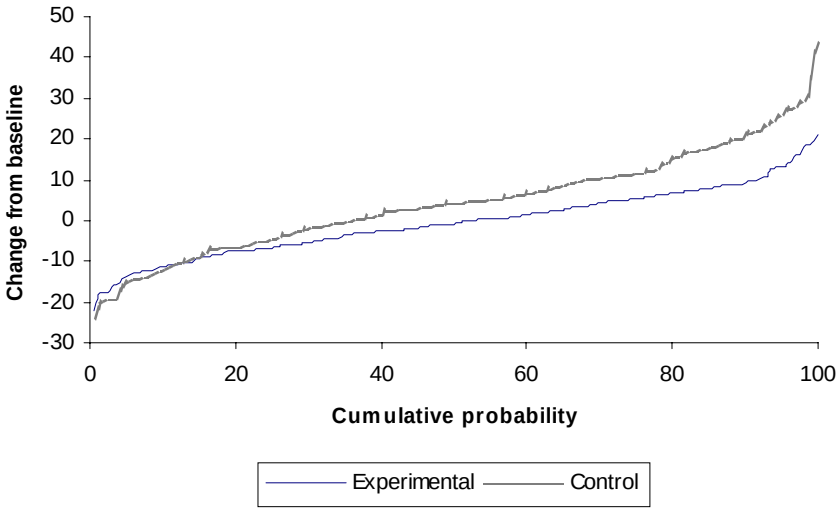

Table 1. Possible combinations of imputation strategies and data analysis

Approach	Imputation strategies	Data analysis
1	Case complete analysis	T test
2	Case complete analysis*	Linear mixed effect model
3	LOCF	T test
4	LOCF	Linear mixed effect model
5	Multiple imputation	T Test
6	Multiple imputation	Linear mixed effect model

*All available data in the context of linear mixed effect model

Table 2. Results of simulation study for missing value around 11% (a) and 22% (b) at 2 years.

Data Analysis	Imputation strategies	Alpha	Treatment effect			Standard deviation	
			Power (%)	MB	RMSE	MB	RMSE
(a)							
	Complete data set						
	Case complete						
T test	LOCF						
	Multiple imputation						
	Complete data set						
Linear mixed	All available data						
effect model	LOCF						
	Multiple imputation						
(b)							
	Complete data set	4.6	79.0	-0.3	37.0	+34.1	34.5
	Case complete	8.1	50.3	-19.3	47.6	+34.1	34.6
T test	LOCF	6.0	58.5	-17.8	43.4	+30.7	31.1
	Multiple imputation	7.2	55.1	-15.2	46.2	+50.1	51.5
	Complete data set	4.5	78.7	-0.3	37.2	-0.1	7.3
Linear mixed	All available data	6.2	58.0	-13.3	44.6	-2.0	8.4
effect model	LOCF	6.7	58.6	-17.9	43.2	-1.6	7.9
	Multiple imputation		56.0	-14.0	46.0	-0.3	8.6