PharmaSUG 2015 - Paper SP03

MISSING DATA FOR REPEATED MEASURES: SINGLE IMPUTATION VS MULTIPLE IMPUTATION

Giulia Tonini, PhD Menarini Ricerche, Florence, Italy

Simona Scartoni, Menarini Ricerche, Florence, Italy

Camilla Paoli, Menarini Ricerche, Florence, Italy

Andrea Nizzardo, Menarini Ricerche, Florence, Italy

Angela Capriati, MD, PhD, Menarini Ricerche, Florence, Italy

ABSTRACT

Missing data is often a major issue in clinical trials, especially when the outcome variables come from repeated assessments. Single imputation methods are widely used. In particular, when data collection is interrupted at a certain time point, Last Observation Carried Forward (LOCF) is usually applied. Regulatory agencies advise to use the most conservative approach to impute missing data. As a drawback, single imputation methods do not take into account imputation variability.

In this work we intend to compare single imputation versus multiple imputation methods in order to verify the effect on the successive inferential analysis, especially in terms of statistical significance of results. In particular we intend to verify if a more conservative single imputation method can be considered also a conservative method in term of statistical significance, respect to multiple imputation, where the higher variability can reduce the probability of having a significant result.

We simulated a dataset representing a clinical trial testing the analgesic efficacy of a combination of drugs on moderate to severe pain after surgery. Pain is measured using a VAS scale. Analysis of covariance is applied to the primary efficacy variable, which is VAS change versus baseline.

Both methods for handling missing data are applied. Multiple imputation in SAS uses PROC MI. We finally present statistical significant results. Analyzing results from several simulated dataset, we found out that multiple imputation consistently reduce the probability of finding statistical significance.

INTRODUCTION

Missing data are part of almost all clinical trial and the decision on how to deal with it is crucial in the outcome of the trial itself. It is in fact inevitable that some patients will drop-out before completing the trial, either by discontinuing their prescribed course of treatment (non-compliance) or by ceasing to be evaluated (drop-out) or both. In this paper we consider drop-outs. In general, data from such trials can be analyzed in different ways: discard data from all patients who did not complete the trial and analyze the remaining data, analyze only the observed data, use a single or multiple imputation to replace the missing observation with plausible values, then analyze the completed dataset.

In this paper we consider drop outs in longitudinal data. In order to avoid biased results, it is important to consider the mechanism of missingness. A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is called *non-random* (MNAR). In the context of likelihood inference and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are ignorable, while a non-random process is non-ignorable.

In many clinical trials, the standard methodology used to analyze incomplete longitudinal data is based on such methods like *last observation carried forward* (LOCF), *complete case analysis* (CC) or simple forms of imputation. This is often done without questioning the possible influence of these assumptions on the final results.

Numerous missing data methods are formulated as selection models (Little and Rubin, 1987), where the joint distribution of the measurement and response mechanisms is given by the marginal distribution, conditional on the measurements. In this way it is straightforward to make inference about the treatment effects and its evolution over time.

Mallinckrodt et al (2003) and Lavori et al (1995) propose direct-likelihood and multiple-imputation methods to deal with incomplete data. Siddiqui and Ali (1998) compare direct-likelihood and LOCF methods. Molenberghs, Thijs, Jansen, Beunckens et al (2004) compare the effects of different methods on three different clinical trials.

This paper will focus on a case study where repeated measures are used to evaluate the analgesic efficacy of oral Desketoprophen and Tramadol fixed combination on moderate to severe pain after elective hip arthroplasty.

The treatment and assessment period was 5 days long. During these 5 days patients were hospitalized. During the first single dose phase, patients received one dose of treatment medication during the first 8 hours, while during the multiple dose phase, a total of 12 doses were administered. Patients continue recording their pain at several time points during both phases.

The reason why some assessments might be missing is due to the fact that the patient could be asleep or unavailable for other reasons. Moreover, an imputation method was also applied in case of use of rescue medication, in order to minimize the impact of RM use on efficacy assessments.

The efficacy evaluation was based on the patient's assessments recorded on an e-Diary, of pain intensity at rest and pain relief (PAR) at pre-defined intervals, patient global evaluation (PGE), worst pain during movements, use of RM and discontinuation from the study for any cause. The primary variable of interest was the mean sum of pain intensity differences (SPID) at rest. The assessments were immediately prior to any study drug intake and at 2, 4 and 6 hours post-dosing with the last assessment prior to the afternoon intake on day 3.

THE CASE STUDY

The study is a phase III, multicentre, randomized, double-blind, double-dummy, parallel-group, placebo and active-controlled study that will be conducted in approximately 40 European and Asian-Pacific sites. 641 male and female patients aged 18 to 80 years, scheduled to undergo standard primary (first-time) one-sided total hip replacement surgery due to primary osteoarthritis, requiring hospitalization for at least 5 days after surgery, were recruited to participate in this study.

After cessation of post-operative analgesic care (1 hour or 2 hours must elapse in case of i.v. or i.m. administration, respectively) and as soon as patients are capable of swallowing oral medications, patients who experience pain at rest of at least moderate intensity (PI-VAS $\geq$ 40mm) the day after surgery are eligible to progress with randomization. Eligible patients were randomized in a 3:3:3:1:1:1 ratio to one of the six possible treatment arms

The treatment and assessment period will encompass two study phases, a single-dose phase (first 8 hours) and a multiple-dose phase starting after the single-dose phase.The multiple-dose phase will start with the second dose administration when the 8 hour single-dose phase is completed. The study medication will be administered every 8 hours for a maximum of 12 doses. This phase will finish 8 hours after the last study drug administration, when the assessment procedures are completed (afternoon of day 5).

The efficacy evaluation will be based on the patient's assessments recorded on an e-Diary, of PI at rest and pain relief (PAR) at pre-defined intervals, patient global evaluation (PGE), worst pain during movements, use of RM and discontinuation from the study for any cause.

The primary objective of the study was to evaluate the analgesic efficacy of oral DKP.TRIS and TRAM.HCl fixed combination on moderate to severe pain after elective hip arthroplasty.

The results of the primary analysis ($SPID_8$) confirmed the superiority of DKP.TRIS + TRAM.HCl over DKP.TRIS monotherapy (p=0.019), and TRAM.HCl monotherapy (p=0.012). In addition, the comparisons of TRAM.HCl and DKP.TRIS versus placebo were both statistically significant ($p<0.05$), thus confirming the model sensitivity.

The present work focus on the multiple-dose phase only (from day 1 dose 2 on). All analyses here reported are performed on the measurements of PI-VAS at each scheduled time point and the change from baseline, that represents the difference of pain that patients have reported in the diary at baseline. The aim of the analyses is to assess the superiority of the combination of Desketoprophen and Tramadol versus the two single components.

In Table 1 the percentages of missing data for each time point is shown.

APPROACHES FOR HANDLING MISSING DATA

FIRST APPROACH

For the ITT population primary analysis, single missing values between measurements will be linearly interpolated. However if more than one consecutive data point is missing the last observation carried forward (LOCF) approach will be used to impute missing data. Using the LOCF approach would see a patient's condition remain relatively constant for the duration of the study in the absence of treatment, and therefore use of the LOCF approach is unlikely to produce a biased estimate in favor of DKP.TRIS + TRAM.HCl, DKP.TRIS or TRAM.HCl. When a data point is missing, the diary will ask for a reason. If a missing value is due to the patient being asleep, this will be replaced with the lowest PI-VAS recorded in the relevant 8 hour period. In cases of consecutive missing values, this rule will be applied only for the last missed value. The others will be handled using a LOCF approach.

Some examples showing possible scenarios are:

1. If there is one missing value which is not due to the patient being asleep between measurements, this value will be linearly interpolated:

PI-VAS	Time Point 1	Time Point 2	Reason for missed value	Time Point 3	Time Point 4
Actual value	60	Missing Value	Different from sleeping	70	73
Replacement value	60	**65**		70	73

2. If there is more than one consecutive missing value, none of which are due to patients being asleep, then the LOCF approach will be applied:

PI-VAS	Time Point 1	Time Point 2	Time Point 3	Reason for missed value	Time Point 4
Actual value	60	Missing Value	Missing Value	Different from sleeping	73
Replacement value	60	**60**	**60**		73

3. If there is one missing value due to the patient being asleep between measurements, this value will be replaced with the lowest PI-VAS recorded in the relevant 8 hours:

PI-VAS	Time Point 1	Time Point 2	Reason for missed value	Time Point 3	Time Point 4
Actual value	73	Missing Value	sleeping	70	60
Replacement value	73	**60**		70	60

4. If there is more than one consecutive missing value due to the patient being asleep, the last missing value will be replaced with the lowest PI-VAS recorded in the relevant 8 hours. For the other missed assessments, a LOCF approach will be applied:

PI-VAS	Time Point 1	Time Point 2	Time Point 3	Reason for missed value	Time Point 4
Actual value	73	Missing Value	Missing Value	sleeping	60
Replacement value	73	**73**	**60**		60

In order to minimize the impact of RM (or paracetamol for antipyretic use) on efficacy assessments, if RM is taken during the single-dose phase then the following rules will apply:

- If RM is taken during the first two hours of the single-dose phase [T(0) – T(2)] then PI should be returned to its baseline level and PAR to zero (so baseline observation carried forward, (BOCF) will be applied) for all subsequent timepoints in the single-dose phase.
- If RM is taken during the last 6 hours of the single-dose phase [T(2) – T(8)] then PI should be returned to its baseline level and PAR to zero for the six hours after intake of RM. If these six hours include timepoints in the multiple-dose phase, then these are also returned to the baseline value but only if at least one value is collected for that patient in the multiple-dose phase. If no values are collected in the multiple-dose phase, then only timepoints in the single-dose phase will be returned to the baseline value.

If RM/paracetamol is taken during the multiple-dose phase, PI recorded during the 6 hours after the intake of RM/paracetamol will be replaced with the LOCF or worst observation carried forward (WOCF) in case the assessment immediately before intake is missed.

SECOND APPROACH

Multiple Imputation method has received a significant amount of attention in recent literature. The idea of multiple imputation first proposed by Rubin [Rubin, 1978], is to impute more than one value for the missing item. The advantage of multiple imputation is that it represents the uncertainty about which value to impute. This is opposed to the first approach which can lead to an underestimation of the variability [Rubin, 1991]. Multiple imputations can be implemented for either longitudinal measurements or a single response. The general strategy is to replace each missing data with a certain number of plausible values from an appropriate distribution. Imputing m values for each missing item, we obtain m different complete datasets. On each of these dataset the planned statistical analysis is applied. Results are then combined so that the final inferential result takes into account the uncertainty caused by the missing data, estimated from the variability of the m independent imputations. If Q_i is the estimate of the unknown parameter and v_i is the variance associated to the i-th imputation, then the final estimate for Q is the mean of the m different estimates

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m} Q_i$$

The variance of$\bar{Q}$ is the sum of a component within imputation variability and a component between imputations variability. Variance within imputation is $\bar{v} = \frac{1}{m}\sum_{i=1}^{m} v_i$ while between imputations variability is

$$B = \frac{1}{m-1}\sum_{i=1}^{m} (Q_i - \bar{Q})^2$$

Total variance associated to $\overline{Q}$ is then $V = \bar{v} + \left(1 + \frac{1}{m}\right) B$

The process is applied as follows:

•We impute m values for the missing item (obtaining m different data sets)

• We replace each missing value with more than one value from an appropriate distribution.

• We perform m different analyses on the m imputed datasets.

Results are then combined so that the final estimate takes into account uncertainty coming from missing data, which is estimated from the variability of the m independent outputs. The parameter of interest is estimated as well as its variance.

The number of imputations needed depends on the amount of missing information. Little and Schenker [Little et al., 1995] and Rubin [Rubin et al. 1991] indicate that 5 imputations are sufficient when the portion of missing data is around 30%.

In SAS, proc MI , together with proc MIANALYZE, can implement the previously described multiple imputation process. It uses methods that incorporate appropriate variability across the m imputations. The method of choice depends on the patterns of missingness.

The most popular method used by PROC MI is the Monte Carlo Markov Chain (MCMC) algorithm. This is based on the assumption of multivariate normality [Schafer, 1997] which implies that valid imputations may be generated by linear regression equations. The algorithm can be applied for both monotone or non-monotone

missing data patterns.

APPLICATION TO THE CASE STUDY

A dummy randomization list was created in order to reproduce the effect of two simple analgesics and a third resulting from the combination of the two.
The model used was the mixed effect model for repeated measures (PROC MIXED). In Fig 1 it is clearly shown how the combination produces a larger decrease in pain respect to the single components. The analysis is performed considering:

- All measurements for each dose
- Only time point 0h and 4h for each dose
- Only time point 0h for each dose

The analysis performed on all measurements gives statistically significant results, as we can see in Table 2. We considered alpha equal to 0.05 since our aim is to assess the significance of the co-primary endpoint: efficacy of the combination respect to both the single agents. In Table 2 treatment differences, 95% CI and p-value are reported.
Considering time point 0h and 4h only, the analysis gives the results showed in Table 3. The combination is still statistically significant respect to both single treatments.
We then applied multiple imputation. The method uses PROC MI in SAS and performs 5 imputations. The data has a non-monotone missing pattern and the variables considered are all continuous. This means that a multivariate normal distribution can be used. The dependent variable to be imputed is change versus baseline for each time point, while independent variables are baseline, treatment, age and gender. The covariates added in the model has been chosen as they might be predictive of the perception of pain; for this reason, they are used in the generation of the imputed values.
Results are reported in Table 4, when all measurements are considered. We can see that there is a large increase of the standard errors and that the difference among combination and single agents is no more statistically significant. In Table 5 we show the results for the 2 time points case, where the combination demonstrate a statistically significant superiority.
We run both analyses on the dataset with the real randomization list as well. We considered data from the multiple phase (all time points). The treatment effect was strongly statistically significant in both cases of imputation using MI and per protocol imputation (pval=0.008 and pval<0.001 respectively), even if the pvalue was slightly larger when using PROC MI.

SIMULATIONS

We simulated a dataset where two different arms are followed up for three days in order to have repeated assessments. The dataset is created with random missing data. To each simulated dataset both LOCF and MI are applied to impute missing data. Then the dataset is analyzed with a PROC MIXED.
The simulated datasets are created in order to have similar characteristics of the case study dataset: three days of treatment with three doses per day and four time points for the assessments. Mean value and standard deviation for the pain baseline values are taken from real data.
We run two different type of simulations. In the first one we created datasets with an effect related to treatment, while in the second we sampled random dataset with no effect.
We considered different sample sizes: 60 patients, 200 patients, 600 patients.

RESULTS

Considering a sample size of 60 patients only (30 per arm) and with an effect size larger than 30%, over 100000 simulations, we obtained a statistically significant result in 98% of cases, when the missing data were imputated using LOCF, while we got 96% of statistically significant results when multiple imputation was applied. Considering larger sample sizes, the power of the analysis quickly increase toward 100%.
In a second set of simulations, we added no effect related to treatment. This means that the values for pain where completely random and that the null hypothesis was true. We applied PROC MIXED considering the treatment as an independent variable and tested for its significance calculating F statistics. Running 100000 simulations and applying the LOCF method we obtained the type I error rate varying the sample size. Results are shown in Fig 2. We start with a sample size of 60 patients to find a type I error of 14,5%: this means that in the 14,5% of cases we found a statistically significant result while it was not. Using multiple imputation the type I error rate was reduced to 3,6% of cases. Increasing the sample size, we see that using LOCF, the type I error converges toward 5%, while using MI it remains always under 5% as it is at the beginning.

DISCUSSION

Applying the two different method for handling missing data to our case study, where repeated measures were performed, we found slightly different results in terms of statistical significance. While LOCF or similar approaches, can be considered more conservative, we found that with MI the statistical significance might be reduced. This difference has to be addressed to the fact that, even less conservative, MI takes into account uncertainty of the imputation procedure.

To better analyze the situation, we performed simulations. We found no difference in the power of the analysis. Testing the methods when the null hypothesis is false gave satisfying results in any case. When considering the case where the null hypothesis is true, we found much more satisfying results with MI, especially when the sample size is small. LOCF, as it can be seen in Fig 2, strongly depends from the sample size. The type I error rate is smaller than 5% only for sample size over 1600 patients.

This simulation study shows that a less conservative method, which take into account imputation uncertainty, is to be preferred, especially for smaller sample sizes.

CONCLUSIONS

The aim of this paper was to better investigate the difference in two methods for handling missing data in clinical trials with repeated measures. While standard methods like LOCF are more frequently applied, multiple imputation started to be commonly used as sensitivity analysis in clinical trial. The great improvement in using MI is that the method take into account imputation uncertainty. The case study showed an interesting results since the more conservative method showed to have more statistically significant results respect to multiple imputation. To better investigate the problem a simulation study has been performed. In particular we analyzed two different case: - false null hypothesis – true null hypothesis. We found that when there is an effect (null hypothesis is false) both methods have a satisfactory performance in terms of power of the analysis. When there is no effect (null hypothesis is true) we found that there is a large difference when the sample size is small. The dataset where multiple imputation method was applied showed a much lower type I error rate respect to the dataset where LOCF was applied, despite of the fact that LOCF is a more conservative method. We than repeated the analysis with larger sample sizes. We found that the type I error rate in the case where LOCF is applied is strongly reduced as the sample size increase. Further investigations will be focused on analyzing what happens in term of type I error rate, varying the number of repeated measures, together with the sample size.

The results of this paper indicate that when the sample size is limited, multiple imputation method has to be preferred respect to other method like LOCF which are considered more conservative. A sensitivity analysis using MI should always be included in the analysis of a clinical trial.

Figure 1. Means of change from baseline for 0h-2h- 4h-6h time points (no correction for missing)

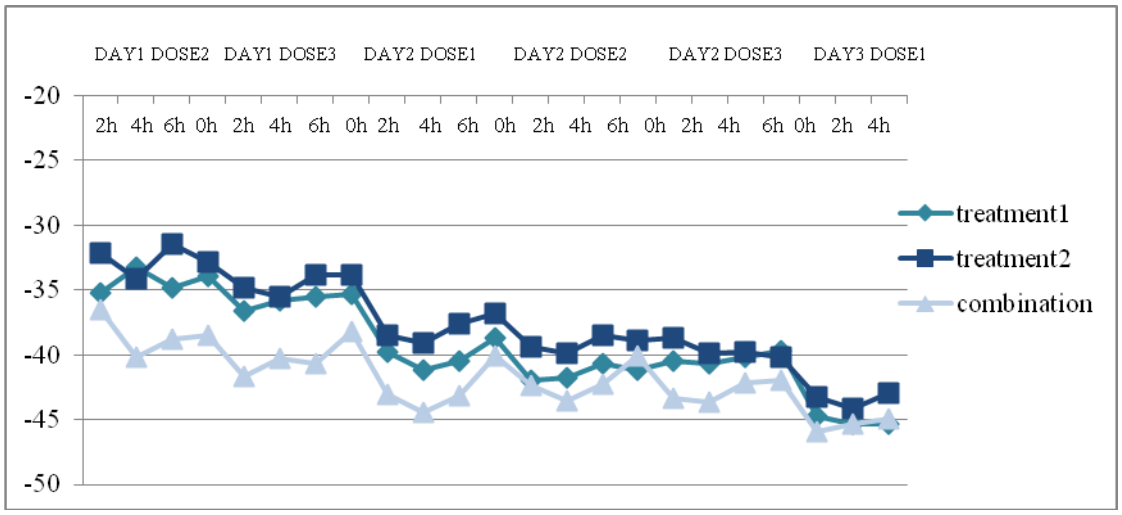

Figure 2. Type I error from simulations varying sample size

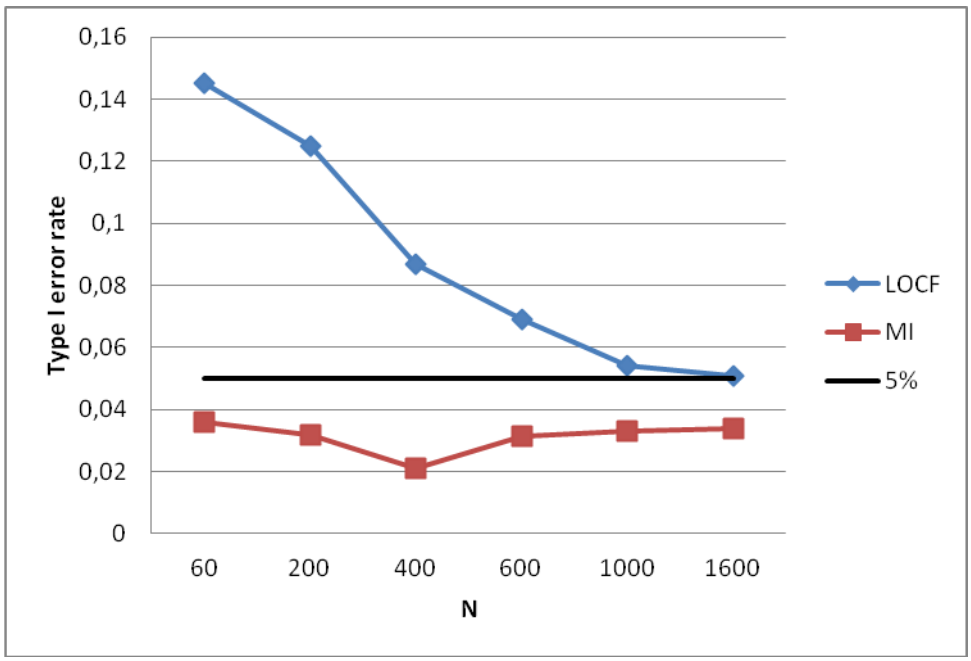

Table 1.

	Desketoprophen	Tramadol	Combination
DAY1DOSE2_2h	7.54%	9.72%	12.67%
DAY1DOSE2_4h	9.43%	7.40%	10.79%
DAY1DOSE2_6h	16.03%	11.57%	13.14%
DAY1DOSE3_0h	14.62%	10.64%	10.79%
DAY1DOSE3_2h	28.77%	28.70%	28.63%
DAY1DOSE3_4h	29.16%	26.85%	25.82%
DAY1DOSE3_6h	20.75%	19.90%	22.53%
DAY2DOSE1_0h	13.67%	10.18%	13.14%
DAY2DOSE1_2h	10.84%	8.79%	12.20%
DAY2DOSE1_4h	12.26%	11.57%	8.92%
DAY2DOSE1_6h	10.84%	12.96%	8.45%
DAY2DOSE2_0h	13.67%	12.5%	8.92%
DAY2DOSE2_2h	13.20%	10.64%	10.32%
DAY2DOSE2_4h	13.20%	11.11%	9.38%
DAY2DOSE2_6h	14.62%	15.27%	11.26%
DAY2DOSE3_0h	15.09%	14.35%	11.73%
DAY2DOSE3_2h	31.60%	26.38%	24.41%
DAY2DOSE3_4h	34.90%	28.24%	25.35%
DAY2DOSE3_6h	23.11%	22.68%	20.65%
DAY3DOSE1_0h	9.90%	14.81%	13.61%
DAY3DOSE1_2h	12.73%	12.96%	8.92%
DAY3DOSE1_4h	12.26%	12.5%	11.26%
DAY3DOSE1_6h	10.37%	12.96%	8.45%
Total	16.46%	15.33%	14.41%

Table 2.

Label	Treatment coefficient estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Combination vs Tramadol	5.0572	1.6791	636	3.01	0.0027	0.05	1.7599	8.3545
Combination vs Desketop.	3.9767	1.6908	636	2.35	0.019	0.05	0.6565	7.2968

Table 3.

Label	Treatment coefficient estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Combination vs Tramadol	5.8716	1.7332	636	3.39	0.0007	0.05	2.468	9.2751
Combination vs Desketop.	5.4965	1.7454	636	3.15	0.0017	0.05	2.0691	8.9238

Table 4.

Label	Treatment coefficient estimate	Std Error	LCL Mean	UCL Mean	Min	Max	t for H0: Parameter= Theta0	Pr > |t|
Combination vs Tramadol	2.81	1.6	-0.34	5.98	2.48	3.46	1.75	0.081
Combination vs Desketop.	2.34	1.6	-0.94	5.64	1.43	2.79	1.41	0.1614

Table 5.

Label	Treatment coefficient estimate	Std Error	LCL Mean	UCL Mean	Min	Max	t for H0: Parameter=Theta0	Pr > |t|
Combination vs Tramadol	3.57	1.6	0.25	6.89	3.29	4.44	2.11	0.0351
Combination vs Desketop.	3.75	1.6	0.42	7.08	3.11	4.43	2.22	0.027

REFERENCES

Horton, NJ, Lipsitz SR, Parzen M, (2003) A potential for bias when rounding in multiple imputation. *American Statistician* 57: 229-232.

Lavori, P.W., Dawson, R., and Shera, D. (1995), A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data, *Statistics in Medicine*, 14, 1913–1925.

Little RJA, Schenker N. Handbook of statistical methodology – Missing data. *New York: Plenum Press*; 1995.

Rubin DB, Multiple imputations in sample surveys – A phenomenological Bayesian approach to nonresponse. Imputations and editing of faulty or missing survey data. *U.S. Department of Commerce*. 1978:1-23

Rubin DB, Schenker N. Multiple imputation in health care databases: an overview and some applications. *Statistics in Medicine* 1991;91:473-489.

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys, *New York: John Wiley & Sons, Inc*.

Schafer, J.L. (1997), Analysis of Incomplete Multivariate Data, *New York: Chapman and Hall*.

Menarini Group. Oral Treatment for Orthopaedic Post-operative Pain With Dexketoprofen Trometamol and Tramadol Hydrochloride (DAVID-art). In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2013 Jul 15]. Available from: https://clinicaltrials.gov/ct2/show/NCT01902134 NLM Identifier: NCT01902134.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Giulia Tonini

Menarini Ricerche

Via Sette Santi 1

50131 Firenze, Italy

Email:gtonini@menarini-ricerche.it