Nutrición Hospitalaria
ISSN: 0212-1611
info@nutriciónhospitalaria.com
Grupo Aula Médica
España

Abellana Sangra, Rosa; Farran Codina, Andreu

The identification, impact and management of missing values and outlier data in nutritional epidemiology

Nutrición Hospitalaria, vol. 31, núm. 3, 2015, pp. 189-195
Grupo Aula Médica
Madrid, España

Nutrición Hospitalaria

Nutr Hosp. 2015;31(Supl. 3):189-195
ISSN 0212-1611 • CODEN NUHOEQ
S.V.R. 318

The identification, impact and management of missing values and outlier data in nutritional epidemiology

Rosa Abellana Sangra[1] and Andreu Farran Codina[2]

[1]*Department of Public Health, Faculty of Medicine, University of Barcelona.* [2]*Department of Nutrition and Food Science, Faculty of Pharmacy, University of Barcelona. Spain.*

Abstract

When performing nutritional epidemiology studies, missing values and outliers inevitably appear. Missing values appear, for example, because of the difficulty in collecting data in dietary surveys, leading to a lack of data on the amounts of foods consumed or a poor description of these foods. Inadequate treatment during the data processing stage can create biases and loss of accuracy and, consequently, misinterpretation of the results. The objective of this article is to provide some recommendations about the treatment of missing and outlier data, and orientation regarding existing software for the determination of sample sizes and for performing statistical analysis. Some recommendations about data collection are provided as an important previous step in any nutritional research. We discuss methods used for dealing with missing values, especially the case deletion method, simple imputation and multiple imputation, with indications and examples. Identification, impact on statistical analysis and options available for adequate treatment of outlier values are explained, including some illustrative examples. Finally, the current software that totally or partially addresses the questions treated is mentioned, especially the free software available.

(*Nutr Hosp* 2015;31(Supl. 3):189-195)

DOI:10.3305/nh.2015.31.sup3.8766

Key words: *Missing data. Outliers. Data collection. Epidemiology nutritional.*

Correspondence: Rosa Abellana Sangrà.
Facultad de Medicina. Universidad de Barcelona.
C/ Casanova, 143.
08036 Barcelona.
E-mail: rabellana@ub.edu

IDENTIFICACIÓN, IMPACTO Y MANEJO DE LOS VALORES AUSENTES Y DATOS ATÍPICOS EN EPIDEMIOLOGÍA NUTRICIONAL

Resumen

Cuando se realiza un estudio epidemiológico nutricional, es inevitable que aparezcan valores ausentes y atípicos. Los datos ausentes aparecen, por ejemplo, por la dificultad de recoger los datos en las encuestas dietéticas que conducen a una falta de información sobre la cantidad de alimentos consumidos y una pobre descripción de ellos. Un inadecuado tratamiento durante el proceso de recolección nos conduce a sesgos y pérdida de precisión y consecuentemente una incorrecta interpretación de los resultados. El objetivo de este artículo es proporcionar recomendaciones sobre el tratamiento de datos ausentes y atípicos, y algunas orientaciones sobre el software existente para calcular el tamaño de muestra y realizar el análisis estadístico. También se realizan recomendaciones sobre la recolección de datos que es un paso importante en la investigación nutricional. Se comentan los métodos que se usan para hacer frente a los datos ausentes, específicamente, el método eliminación de casos, imputación simple o múltiple con indicaciones y ejemplos. También se relata cómo se identifican datos atípicos, el impacto que tienen en el análisis estadístico, las opciones para un adecuado tratamiento y se ilustra mediante un ejemplo. Finalmente, se menciona el software existente que aborda total o parcialmente las cuestiones tratadas, específicamente el software de libre distribución.

(*Nutr Hosp* 2015;31(Supl. 3):189-195)

DOI:10.3305/nh.2015.31.sup3.8766

Palabras clave: *Valores ausentes. Valores atípicos. Recogida de datos. Epidemiologia nutricional.*

Abbreviations

WHO: World Health Organization.
MCAR: Missing Completely at Random.
MAR: Missing at Random.
NMAR: Not Missing at Random.
BMI: Body mass index.
EM algorithm: Expectation Maximization algorithm.
GUI: Graphical user interface.

Introduction

According to the World Health Organization (WHO), epidemiology is "the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems". Nutritional epidemiology is focused on the aspects of diet that can influence the occurrence of human disease.

Diet is a complex set of exposures that are strongly intercorrelated. Individuals are exposed to diet in different grades with very few clear changes in diet at identifiable points in time. Assessment of food intake is difficult and is subjected to multiple biases. Additionally, the consumption of nutrients is usually determined indirectly based on the reported food consumed or on the level of biochemical measurements. Thus, the most serious limitation of research in nutritional epidemiology is the measurement of exposures to dietary and nutritional factors[1].

Among other problems, missing values in dietary surveys can be due to missing days or meals of intake, missing amounts of foods consumed, inadequate descriptive information for useful coding at an individual level, absence of the consumed food or nutrient of interest in food composition tables, and, in a broader sense, the non-participants in a representative random sample study. If filling in missing gaps is not performed, the effect of this information gap needs to be considered in the interpretation of the results obtained in research studies[2].

The scientific research process can be divided in several stages[3]. Firstly, it is important to review scientific literature and to properly formulate a research objective and related hypothesis. Then, a good research design to answer the question may be elaborated. Sampling procedures and determination of sample size are important parts of this design. All this information should be explained in a research protocol, in which instrumental and procedural details of the study must be stated; for example, questionnaires, biochemical analyses or other procedures that will be used in data collection. Validation of food intake questionnaires is an important point to avoid biases in data. Once the protocol is finished and has been reviewed, field work and data collection can be started. These data obtained should be coded and processed, and this data process is an important part in dietary surveys, especially if food consumptions are used to estimate nutrient intakes. Adequate software allows large data volumes to be entered, managed and processed for the purpose of preparing a data matrix for statistical analysis. In this step, the identification and extraction of outlier values and the completion of missing data are essential to avoid troubles in statistical analysis. Different methods have been developed to deal with outliers and missing data, and correct choice of these methods is critical.

The purpose of this article is to provide some recommendations about the treatment of missing and outlier data, and orientation regarding existing software for the calculation of sample sizes and performing statistical analysis. This software can help to prevent possible incorrect results in statistical analysis and misinterpretation of the observations made.

Data collection. Recommendations.

Data collection is an important part of research. The information of the subjects to be recorded and how the variables are measured conditions the posterior statistical analysis and the validity of the study. It is recommended that original variables be recorded instead of calculated or categorical variables used. For example, the year of birth should be asked instead of the age of the subject or the nutritional status should not be reported with the categories; low weight, normal, overweight and obesity. It is better to ask for the weight and health of the subject and later compute the body mass index. The same occurs with food frequency intake. If the variable frequency is recorded using the categories less than twice a week, between 2 and 5 times and 5 or more times a week. It is impossible to know the number of intakes of a particular food and thus, the variable cannot be modified.

The categorical variables should be correctly codified with their appropriate labels. Usually a number is assigned to each category and then the label is given. This is advisable because most software are key sensitive or detect the accent of the words and a new category is falsely created. This can be solved but not after having lost considerable time.

In the case of recording data using specific software the introduction of validation rules that restrict the range values of the variables is recommended to avoid errors of digitalization.

Missing data and Outliers

Missing data

Once the data have been recorded it is important to consider the information not provided by the subjects, that is, the missing data. According to Rubin (1976)[4] there are basically three types of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). The first refers to the probability of an observation being missing does not depend on observed or unobserved measurements. Missing at random refers to the fact that the probability of an observation being missing depends only on the values of the observed study variables, and not missing at random refers to the probability that an observation is missing depends on unobserved measurements.

For example, in the supposition that the body mass index is recorded according to the sex of the subjects, then, regardless of any particular reason why some subjects disclose their weight, the missing data are considered missing completely at random. However, if females are likely to not reveal their weight, missing data depends on the sex and these data are considered to be missing at random. And finally, in the case that obese subjects are more likely to disclose their weight, the probability that body mass index is missing depends on the unobserved value itself. These data are not missing at random.

Illustrative data

A random sample of 58 students of Human Nutrition and Dietetics and Food Science and Technology of the University of Barcelona were studied to evaluate their nutritional status. The food consumption of the students was recorded using the 24h recall, 3-day recalls and the food frequency questionnaires. The students also filled in a questionnaire for assessment and quantification of overweight and obesity-related lifestyles[5]. The aim was to study the relationship between protein intake with respect to sex, body mass index (BMI), total energy intake and index of overweight and obesity. Eight female students did not report their weight and height, thus the BMI had 8 missing data (Table I).

Case deletion

There are two ways by which you can delete cases that have missing values; listwise and pairwise deletion. The listwise method deletes the subjects with missing data. If the data are missing completely at random, listwise deletion does not add any bias, but the sample size is reduced thereby affecting the power of the tests (decreasing) or the standard errors of the estimations (increasing). Moreover, listwise deletion discards considerable information provided by the subject.

Pairwise deletion (or "available case analysis") deletes a subject when data are missing in a variable required for a particular analysis, but includes that subject in analyses for which all required information are present. When pairwise deletion is used, the total sample size for analysis will not be consistent across parameter estimations.

Table II shows the estimates of the linear regression of logarithm of proteins according to sex, BMI, obesity index and total energy intake. The information of the students with missing data in BMI variable is deleting (listwise deletion).

Simple Imputation or multiple imputations

Imputation is the process of replacing missing data with substituted values. Several methods are used: mean imputation, regression imputation, expectation maximization algorithm and multiple imputations.

Mean imputation consists in replacing any missing data by the mean of non-missing data. In our data all the missing data are replaced by the mean BMI (21.32).The problem of the mean imputation is that it attenuates any correlations between the variables that

Table I
Data from 12 students. Missing values for variable of BMI are shown as an empty box

Sex	*Obesity Index*	*BMI*	*ENERGY*	*PROT*	*Mean Imputation BMI*	*Regression imputation BMI*
2	80		901.8	53.1	21.33	27.15
2	78		3197.2	177.6	21.33	24.30
2	72		2295.5	96.0	21.33	19.89
2	65		2229.8	113.6	21.33	19.55
2	62		2131.1	79.0	21.33	21.18
2	82		2137.9	125.4	21.33	27.17
2	80		1453.3	69.7	21.33	22.25
2	63		2927.2	124.4	21.33	19.55
2	69	20.10	2684.6	104.8	20.10	20.10
1	58	23.78	2681.9	144.5	23.78	23.78
2	78	22.12	2677.3	136.4	22.12	22.12
1	67	21.29	2674.6	127.5	21.29	21.29
…	…	...	…	…	...	...

BMI: Body mass index, ENERGY: total energy intake, PROT: protein intake

Table II
Estimation of the linear regression of the logarithm of proteins according to sex, BMI, obesity index and total energy intake using the listwise deletion method (left panel) and multiple imputation (right panel)

	Listwise deletion			*Multiple Imputation*			
Variable	*Beta*	*Std. Error*	*P value*	*Beta*	*Std. Error*	*P value*	*R.M.I*
Female	-0.245	0.089	<0.001	-0.240	0.085	<0.001	0.003
BMI	-0.003	0.092	0.009	-0.020	0.013	0.008	0.109
Index obesity	-0.002	0.003	0.001	0.006	0.003	0.001	0.036
Total Energy	0.0003	0.00005	<0.001	0.0003	0.0004	<0.001	0.017
8 observations deleted due to missingness				Non-missing data			

Std. Error: Standard error, R.M.I.: Rate of missing information

are imputed. Table I shows the BMI values using mean imputation.

Regression imputation, the missing data are replaced by the predicted value of the regression derived from the non-missing data. In contrast with the mean imputation, the imputed value is in some way conditional on other information of the subjects. Considering the example of missing data in variable of BMI, with mean imputation all the missing data are replaced by the mean of the BMI. However, with regression imputation you can replace the missing data according to the sex, the overweight and obesity index and total energy intake of the students. The values of BMI filled in by regression imputation are shown in table I. Each student has a different BMI value according to the overweight and obesity index and total energy intake. Thus, there is an improvement on comparing regression imputation with mean imputation, but the predicted value has an error that it is not considered on performing imputation. This causes relationships to be over identified and suggests greater precision in the imputed values than warranted. Nonetheless, this difficulty can be overcome by stochastic regression imputation. This approach adds a randomly sampled residual term from the normal (or other) distribution to each imputed value.

Another way to deal with missing data is a technique called the expectation maximization algorithm abbreviated as the EM algorithm[6]. This method assumes a distribution for the partially missing data and bases inferences on the likelihood under that distribution. It is an iterative process in which the following two steps are repeated until convergence: The E step finds the conditional expectation of the missing data, given the observed values and current estimates of the parameters. These expectations are then substituted for the missing data. In the M step, maximum likelihood estimates of the parameters are computed as though the missing data had been filled in. However, the uncertainty of the missing data are not considered.

In multiple imputation[7] instead of filling in a single value for each missing value, each missing value is replaced by *m* simulated data that represents the uncertainty about the right value to impute. Then, each imputation generates a data set. Each imputed data set is analyzed separately and *m* estimates and their standard errors are obtained. The overall estimate is the average of all the estimates. The standard error of the estimation is performed using the within variance of each data set (average of the *m* standard errors) as well as the variance between the imputed data set (sample variance across the *m* parameter estimates). These two variances are added together and their square root determines the standard error and thus, the uncertainty due to missing data is introduced in the standard error of the estimate. The between variation between imputed data sets also reflects statistical uncertainty due to missing data.

In our data 15 imputations were performed. Thus, we have 15 data sets according to the values of imputation. Table II shows the parameter estimation of the coefficients of the linear regression of logarithm of proteins according to sex, BMI, the students is now used because the missing BMIs are filled in. Furthermore, the rate of missing information quantifies the relative increase in variance due to nonresponse for BMI. The BMI has a rate of 0.109 and the remaining variables are very low because no imputation was performed.

If your data are MCAR, then there is no bias in your data. If only a few cases have missing data, each with many blank spaces, the best option is to choose listwise deletion. If your data are MAR, the best solution is multiple imputation. Maximum likelihood imputation, and stochastic regression imputation are also suitable, but multiple imputation is recommended. If your data are NMAR these methods are often biased and some specific methods can be used[7].

Outliers

An outlier is an observation clearly different from the rest of the data; it is an atypical or extreme observation. There are several methods to detect outliers; plots such as normal probability plots, box plot or others are model-based.

Model-based detection assumes that the data have a normal distribution, such as the Grubbs' test for outliers[8], Pierce's criterion[9], or Dixon's Q test[10].

One common method for outlier detection is the use of inter-quartile rank. An observation is outlier if the value is outside the limits and ; k is normally set at 1.5 or 3.

It is important to study outlier data because most of the statistics used are influenced by these data and are not robust. For example, the mean is sensitive to the extreme observation, but the median is not. Suppose 10 students have a protein intake of between 50 and 160 g/day but one has an intake of 250 g/day. The mean is 166 g/day while the median is 81 g/day. The mean is affected by this extreme value and the median is not.

In regression analysis the outlier values can influence the results. In regression a difference is made between an outlier and an observation that has high leverage.Concretely, an outlier is an extreme observation in a response variable. However, an observation that has an X value far from its mean is called high leverage. Leverage measures the distance between the mean of the X distribution. When the leverage is two or more times greater than the average leverage , it is considered to have high leverage, being p the number of regression parameters, and *n* the sample size.

Data with high leverage and outliers might have potentially influence in regression. An influential observation generates a negative impact because it biases the estimations. Otherwise, not all the points with high leverage or outliers necessarily influence the regression coefficients. It is possible to have a high leverage and yet follow a straight line with the pattern of the rest of the data.

Figure 1 plots the height and the weight of 60 subjects. The variables have a linear relationship. Three different points have been added to the plot (A, B and C). A is an outlier with respect to the weight variable, but not to height. Its leverage is low (0.016) because it is lower than 2*(2/61)=0.06. B is an outlier with respect to weight and has a high leverage, and C is not an outlier compared to weight but has high leverage.

A preliminary analysis to detect outliers involves the use of residuals of the model. One problem with residuals is that their values depend on the scale and units used. Since the residuals are in units of the dependent variable Y there are no cut-off points for defining what a large residual is. This problem is overcome by using standardized residuals, calculated dividing a residual by its standard error. Observations with absolute standardized residuals exceeding 3 require close consideration as potential outliers. Points A, B and C in figure 1 have a standardized residual equal to 6.39, -0.26 and -5.76. Points A and C have a large residual, however B has a low residual and it is an outlier with high leverage.

The Cook's distance[11] is a measure to detect potentially influential observations. The distance measures the effect of deleting an observation. Data points with large residual (outlier) and/ or high leverage may distort the estimation and accuracy of the regression model. Points with a large Cook's distance, operate under the rule that a value greater than 1 is recommended to study the potential influence??? Another common rule considers a threshold of the percentile 1-alpha of the Fisher Snedecor distribution (F(p,n-p, 1-alpha))

Points A, B and C have a Cook's distance of 0.34, 0.02, and 4.25, respectively. Although A is an outlier it has no potential influential observation, but as shown in table III the standard error of the coefficients increased and the goodness of fit (coefficient of determination) decreased from 0.82 to 0.58 (Table III). Points B

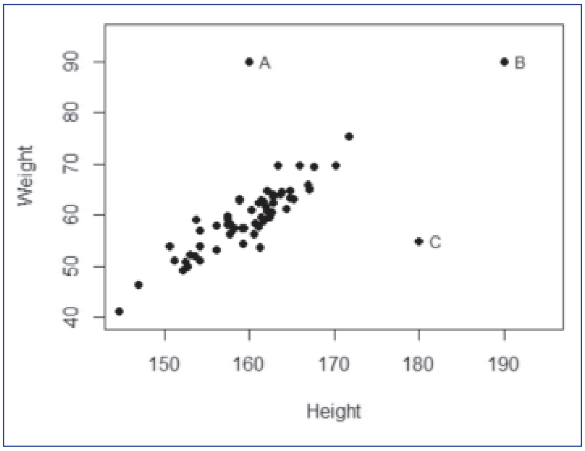

Fig. 1.—Scatterplot of Height and Weight, with three potentially influential observations (A, B and C).

Table III

Estimation of the regression coefficients, standard errors and determination coefficient of the regression of weight and height according to observations (A, B and C) included

	Intercept	*Std. Error Intercept*	*Beta*	*Std. Error Beta*	*R^2*
Without A, B and C observ.	-106.4	10.13	1.04	0.06	0.82
Observation A	-106.8	18.10	1.04	0.11	0.58
Observation B	-104.9	8.17	1.03	0.05	0.87
Observation C	-68.6	13.69	0.80	0.08	0.60

Linear regression: Weight=intercept+ beta*Height, R2 the coefficient of determination.

and C have high leverage but only C has a potential influential observation. Table III shows that the estimates, the standard errors and the coefficients of determination are modified when C is added to the data.

Other diagnostic methods to detect potentially influential observations are: DFBETAS DFFITS and COVRATIO statistics[12]. All measure the impact of a particular observation deleted from analysis, concretely; DFBETAS measures the effect on the estimation of the coefficients, DFFITD on the predicted value and COVRATIO on the variances (standard errors) of the regression coefficients and their covariance.

When an outlier is detected, what arises from the value should first be evaluated. If the value arises from a human or instrument error the mistake should be corrected. However, outlier data may arise from different causes such as the inherent variability of the variable or if the underlying distribution has an asymmetrical distribution or the data is from another population. Alternatively, outliers may suggest that additional explanatory variables need to be brought into the regression analysis. Deletion of outlier data is a controversial practice and rather than omitting them it is recommend to use robust statistical methods which are not excessively affected by outliers.

Statistical software

Sample size software

It is important to compute the number of subjects to study when the objective and the study design have been defined. The sample must be representative of the population studied. According to the aim to the study and the structure of the population, several sampling types can be used: simple random sampling, systematic, stratified or cluster sampling. Furthermore, sample size computation depends on the main objective, the need to have minimum accuracy in the estimation or the need to have sufficient statistical power. It is also important to consider an extra percentage of individuals because there may be missing values. It is difficult to recommend the percentage of individuals and this depends on the area applied. Moreover, it is convenient to design strategies to guarantee or control that the subjects respond to all the information of the questionnaire.

To calculate sample sizes many commercial or free software are currently available. In relation to the free software EPIDAT 4.0 and GRANMO calculate sample size and power according to statistical methodology.

The EPIDAT 4.0 software was created by *Servizo de Epidemioloxía* de la *Dirección Xeral de Innovación e Xestión da Saúde Pública* de la *Consellería de Sanidade (Xunta de Galicia)* with the support of *Organización Panamericana de la Salud and Universidad CES of Columbia*. It can be downloaded from the page http://www.sergas.es/ in the section of Research and Healthcare Innovation/ data/Software.

GRANMO software was developed by the research group of Cardivascular Risk and Nutrition and Cardiovascular Epidemiology and Genetic of the Research Programme on Cardiovascular and Inflammatory Processes of the IMIM - Hospital del Mar. It can be downloaded from the web page http://www.imim.cat in the section transfer opportunities and service offers/ Freeware.

Statistical analysis software

There is a great variety of statistical software available. At present, the free software most commonly used is the R-project[13]. It is a GNU project which was developed at Bell by John Chambers and colleagues. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. R is an interpreted language. R has a command-line interpreter and has several of packages that users can download from the web page. The R project does not have a friendly graphical user interface (GUI) but some packages such as R commander[14] or Deducer[15] provide a menu-driven GUI.

Software which can be purchased include: S plus [16](commercial version of the R project), SPSS[17] (statistical packages for the social sciences), SAS institute[18] (Statistical Analysis System), STATA [19](Statistics and Data) or Minitab[20].

Multiple imputation has become increasingly popular, and all these software implement this technique. Yucel (2011)[21] provides a description of the methodology implemented by software.

Finally, all these software perform a wide variety of statistical analyses and have great power to generate graphs of the results. The most appropriate choice of software therefore depends on the costs and preferences of each user.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Willet W. Nutritional epidemiology. 2nd ed. Oxford: Oxford University Press; 1998.
2. Arab L. Analyses, presentation, and interpretation of results. In: Cameron ME, Van Staveren W, editors. Manual on methodology for food consumption studies. Oxford: Oxford University Press; 1988.
3. Polgar S, Thomas SA. Introducción a la investigación en ciencias de las salud. Madrid: Churchill Livingstone; 1993.
4. Rubin, D.B. Inference and missing data.*Biometrika* 1976; 63: 581-92
5. Pardo A, Ruiz M, Jódar E, Garrido J, De Rosendo J M, Usán L A. Development of a questionnaire for the assessment and

quantification of overweight and obesity related lifestyles. *Nutrición Hospitalaria* 2004;XIX(2):99-109

6. Dempster, AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion) 1977. *Journal of the Royal Statistical Association* 1977; B39:1-38
7. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 1987.
8. Grubbs FE. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics* 1950; 21(1): 27-58
9. Pierce B.Criterion for the Rejection of Doubtful Observations. *Astronomical Journal* II 1852; 45: 161-310.
10. Dean RB and Dixon WJ. Simplified Statistics for Small Numbers of Observations. *Anal Chem*1951; 23 (4): 636–638.
11. Cook R D Influential Observations in Linear Regression. *Journal of the American Statistical Association. 1*979; 74 (365): 169–174.
12. Belsley DA, Kuh E, Welsh RE. Regression diagnostics: identifying influential data and sources of collinearity. Wiley series in probability and mathematical statistics. New York 1980.
13. R-project version 3.1.2. Dowload from: http://www.r-project.org/
14. Fox J, Bouchet-Valat M *et a*l. A platform-independent basic-statistics GUI (graphical user interface) for R, based on the tcltk package. Version 2.1-5.2014
15. Fellows I *et al*. Deducer: A data analysis GUI for R. Version 0.7-7. 2014
16. S- PLUS. TIBCO Software Inc.2014.
17. IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.
18. Statistical Analysis System (SAS) Institute Inc. 2013. Unitate States. Version 9.4.
19. StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.
20. Minitab 17 Statistical Software (2010). [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com)
21. Yucel RM. State of the Multiple Imputation Software. *Journal of statistical software* 2011;45(1).