REVIEW ARTICLE

JORNAL **VASCULAR** BRASILEIRO
ISSN 1677-7301 (Online)

Anomalous values and missing data in clinical and experimental studies

Valores anômalos e dados faltantes em estudos clínicos e experimentais

Hélio Amante Miot[1]

Abstract

During analysis of scientific research data, it is customary to encounter anomalous values or missing data. Anomalous values can be the result of errors of recording, typing, measurement by instruments, or may be true outliers. This review discusses concepts, examples and methods for identifying and dealing with such contingencies. In the case of missing data, techniques for imputation of the values are discussed in, order to avoid exclusion of the research subject, if it is not possible to retrieve information from registration forms or to re-address the participant.

Keywords: data analysis; database; outlier; multiple imputation.

Resumo

Durante a análise dos dados de uma pesquisa científica, é habitual deparar-se com valores anômalos ou dados faltantes. Valores anômalos podem ser resultado de erros de registro, de digitação, de aferição instrumental, ou configurarem verdadeiros *outliers*. Nesta revisão, são discutidos conceitos, exemplos e formas de identificar e de lidar com tais contingências. No caso de dados faltantes, discutem-se técnicas de imputação dos valores para evitar a exclusão do sujeito da pesquisa, caso não seja possível recuperar a informação das fichas de registro ou reabordar o participante.

Palavras-chave: análise de dados; base de dados; discrepância; imputação múltipla.

How to cite: Miot HA. Anomalous values and missing data in clinical and experimental studies. J Vasc Bras. 2019;18: e20190004. https://doi.org/10.1590/1677-5449.190004

[1]Universidade Estadual Paulista – UNESP, Faculdade de Medicina de Botucatu, Departamento de Dermatologia e Radioterapia, Botucatu, SP, Brasil.
Financial support: None.
Conflicts of interest: No conflicts of interest declared concerning the publication of this article.
Submitted: January 08, 2019. Accepted: March 14, 2019.

The study was carried out at Faculdade de Medicina de Botucatu, Universidade Estadual Paulista (UNESP), Botucatu, SP, Brazil.

Before embarking on the process of analyzing the data from a clinical or biomedical study, it is imperative to undertake a careful evaluation of the possibility of missing data or anomalous values in the sample, since they are commonplace and failure to detect them can compromise a study's conclusions or its power of inference.[1] Anomalous values can be the result of errors of recording, of typing, or of readings taken with instruments, or may be true outliers.[2]

As the sample size and/or the number of variables increase, the likelihood of input errors also increases. Studies with very large samples employ techniques such as double-input or review of sub-samples of records, to identify (and prevent) possible errors.

Table 1 shows hypothetical data from a clinical trial in which certain patterns of anomalous values, outliers, and missing data are illustrated.

It can be observed from the sequence of participant identifier numbers that participant number 8 is not included in the records shown in Table 1, which could be because he/she was excluded from the protocol or because of a human input error.

The age column shows one participant's age as 555 years, which is likely to be because a number key has been pressed too many times (for example, 555 instead of. 55 years). However, if the wrong number had been typed and the result is a believable value (such as 23 instead of 32 years, or 4 instead of 44 years), then visual identification of the error would be very much less likely.

Another problem related to recording participants' age is caused by a tendency for research subjects to give their age rounded down to an age younger than their true age (for example, 40 rather than 43 years). In order to minimize this type of bias, it is recommended that participants' year of birth, or even their full date of birth, should be recorded and then their age can be calculated later, when data analysis is conducted. In this case, care should be taken not to record the current date or year instead of the year or date of birth of the participant (for example, 2019 rather than 1979).

Participant 17's sex was recorded as "N", which is a code that is not used for this variable (M or F). Since "N" is the letter adjacent to "M" on the keyboard, this is also a common pattern of input error. Additionally, systems used for statistical analysis may differentiate between higher and lower case letters (for example, "M" and "m") and may also register accents in languages that use them (for example, "não" vs. "nao" in Portuguese). These possibilities can be eliminated by using numerical codes for responses (for example, Male = 1 and Female = 2; Yes = 1 and No = 2).

Sometimes, errors can only be detected by evaluating additional variables, as is the case in record 16, where a participant listed as male reports having had three pregnancies. Along the same lines, record 7 is

Table 1. Example data records (hypothetical) from a clinical study.

Identifier	Age*	Sex	Pregnancies	Systolic blood pressure**	Diastolic blood pressure**	Body mass index#
1	46	F	2	120	80	23.8
2	50	F	3		110	24.9
3	69	F		110	150	22.9
4	22	M	0	135	85	24.1
5	555	M	0	165	95	27.0
6	38		0	125	75	23.9
7	18	F	6	155	90	26.1
9	58	F	3	135	75	24.2
10		M	0	145	85	25.8
11	93	M	0	150	115	24.1
12	45	F	1	135	135	23.7
13	43	F	1	120	80	25.1
14	38		2	140	90	25.0
15	37	M	0	235	180	29.2
16	30	M	3	130	100	24.9
17	42	N	0	120	70	23.8
18	30	F		115	75	
19	25	F	0	135	100	24.2
20	28	M	0	145	105	23.1
21	58	F	3	135	75	24.2

* Age in full years; ** Blood pressure in mmHg; # Body mass index in kg/m^2.

a participant who is only 18 years old, but reports six pregnancies. Finally, participant 21 has exactly the same records as participant 9 for all variables, suggesting double inclusion in the study.

Tests should also be conducted to detect incongruities where values have interdependent behavior. For example, diastolic blood pressure should be lower than systolic, which is not the case in records 3 and 12, in one of which there is a reversal of values and in the other the same value has been input twice.

Errors of measurement caused by incorrectly reading instruments (for example, sphygmomanometers and balances) induce a systemic error that is very unlikely to be detected and corrected. When the error is uniformly propagated throughout the sample (for example, a reading that is 10 mmHg higher for all records), it does not cause such a significant problem for internal comparison of groups. However, when different instruments with calibration problems or poor reproducibility are used, variability is increased and parameters become less exact. Precautions to ensure that data collection instruments or laboratory methods are in agreement are extremely important, because corrections for these biases made during the analytical phase (for example, transforming values into Z scores for the data collected with each instrument) have unsatisfactory performance.[3]

This is an appropriate time to mention that study participants may falsely report some types of information that involve cultural values, for reasons of acceptance, social identification, or moral judgment. In general, values reported for body weight, use of illicit substances, and number of extramarital relationships tend to be underestimated by research participants, whereas reported values for height, use of safe sex methods, and affirmative attitudes (for example, altruism, solidarity, or common sense) tend to exaggerate the true values. There is no infallible method to prevent this type of false report and neither is there any statistical method of correcting for such biases. However, in addition to using objective measures (for example, measuring weight and height during the interview, verifying year of birth on an identity document or hospital records) researchers recommend using confirmatory questions that enable the integrity of information provided to be verified (for example, at the start of the interview ask how many times per month a respondent has used illicit substances and at the end ask how many times a week they use specific substances, marijuana, cocaine, acid, etc.).

The accuracy of records is crucial for a study's quality and the validity of its conclusions; efforts to minimize these types of problems must be considered when planning research.

Data can also contain values that are very different from the behavior of the sample. These are known as outliers and they are not recording errors, but do not fit the probability distribution of extreme values (whether higher or lower) that is found in the population. In the example shown in Table 1, participant 11 is 93 years old, and participant 15 has blood pressure that contrasts with all of the other participants'.

In normal distributions,[4] outlier values are defined as those that are more extreme than 1.5 interquartile deviations below p25 or above p75 in a sample (Figure 1), or standardized values that are beyond three standard deviations (higher or lower) for the sample. Identification of outliers in non-normal distributions, correlation analyses, or multivariate analyses is more complex and is beyond the scope of this review.[5-7]

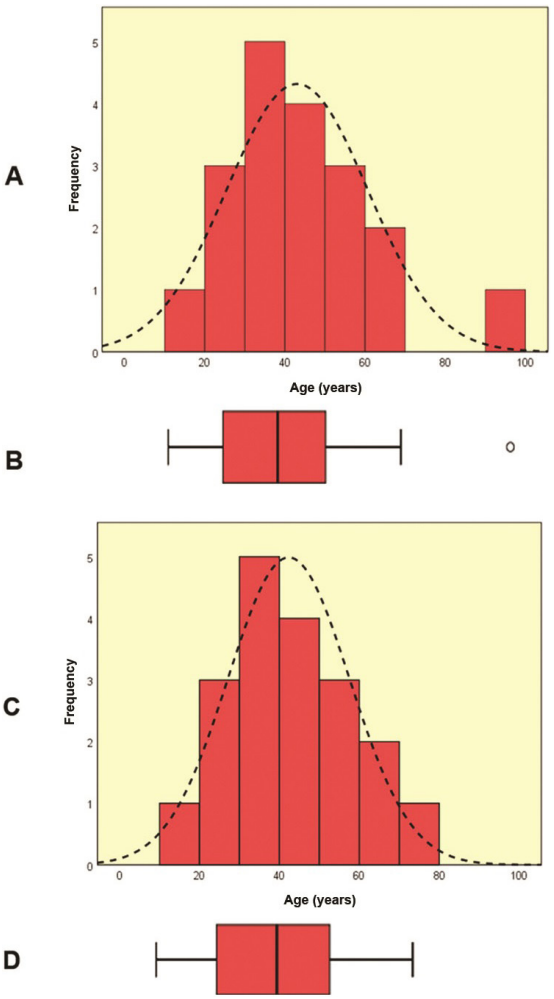

Figure 1. Graphs and box plots for the variable age shown in Table 1, before (A and B) abd after (C and D) winsorization. There was an outlier – the 93 years-old –, which was more extreme than the 1.5 times the interquartile deviation (25 years) added to the 75th percentile (55 years) and was Winsorized to 70 years (n = 19).

However, identification of outliers is just the first step; there is also a great matter of debate on how to deal with these data. If, on one hand, these records are out of tune with the sample, increase the variability of data, compromise the normality of the distribution, reduce statistical power, and have an impact on population inferences, on the other hand, they are real values, from subjects who were part of the study population. Outliers can even be indicative of special patterns within a sample, providing base for new hypotheses on the phenomenon studied, or may reveal underlying non-normal probability distributions in the population.[8]

The bivariate statistical tests habitually used for parametric data (Student's *t* test, ANOVA, and Pearson's correlation coefficients) are relatively robust to deal with a small proportion of outliers. In turn, rank-based tests (Mann-Whitney, Wilcoxon, Kruskal-Wallis, and Spearman's coefficient) are unaffected by extreme values. The decision to exclude subjects with outlier values penalizes the sample, and should be avoided. Rather, if necessary, it is possible to deal with outliers using winsorization, or trimming, or employ clustering techniques, resampling (bootstrapping), or robust statistical analyses, which provide an approximation for a probability distribution, based on the central data.[9-13]

In winsorization, the anomalous datum is substituted with a value that is beyond that of the next nearest value, bringing the outlier closer to the remainder of the data.[1] In the case of Table 1, the age of 93 years could be winsorized to 70 years, one unit higher than the next highest age: 69 years (Figure 1).

In trimming, a certain percentage of the extremes of the sample (for example, the most extreme 2%) is excluded bilaterally from the analysis. This procedure makes the sample more uniform, but it can be at the cost of the power of the statistical analysis, since it reduces the sample size.[1]

Clustering techniques assess patterns of proximity of participants based on the behavior of other variables, and the outlier value is substituted with the average for subjects identified as a group. Clustering techniques, imputation based on resampling methods, and robust statistical methods require the involvement of an experienced statistics professional.[10,11,14-17]

It is important that researchers employ routines for identification of anomalous values and outliers, because of the inferential cost they can impose, especially in studies with small numbers of participants. If outliers occur at low frequencies in the sample and do not change the conclusions of an analysis, it is recommended that data be not transformed in any way.

Another commonplace occurrence in clinical and experimental research is missing data, which can be easily diagnosed visually, by the "space" that they leave on a spreadsheet of data (Table 1). However, as the number of subjects and/or variables increases, it is recommended that strategies to test for missing data be adopted. Furthermore, some spreadsheet and data analysis programs automatically substitute missing data with ZERO or an incorrect value (for example, 999), which can cause even greater problems if these data are not identified.

Missing data may be caused by input errors or they may really have been unavailable when data were collected. If possible, retrieval of original records or returning to the subject for confirmation are the best solutions in these cases. In some cases, the behavior of other variables makes it possible to deduce the missing value with certainty. In Table 1, record 14 must be for a woman, since it shows two pregnancies.[2]

However, some data cannot be recovered *a posteriori* (for example, a patient has died or experimental mice have been euthanized), cannot be deduced, are affected by when they were collected, or are the result of complex experiments. These circumstances demand use of certain statistical techniques to deal with these limitations.[18-20]

The first step in dealing with missing data is to analyze the magnitude of the absence of values. Subjects missing more than 10% of data, or variables with more than 10% of missing values are not suitable for techniques for imputation of values, and retention of the subject or variable in the study should be questioned.

The second step is to analyze patterns in missing data, because techniques for imputation demand that the absence of data is relatively independent of other variables, since the lack of information may itself be linked to the behavior of one of the variables.

Missing data that do not follow any type of pattern of absence are known as missing completely at random (MCAR) data, such as when one sheet of a questionnaire is lost, a single blood sample coagulates, or a patient moves to another town. In such cases, it is assumed that the absences of data are caused by elements external to the protocol, and so analysis of the data with or without those participants with missing data will not change the magnitude of the effect.[21]

There are also missing at random (MAR) data, where the lack of one value is subject to the effect of a secondary covariable: those with less education may leave responses unanswered because they don't understand, questions of a sexual nature may be ignored by promiscuous participants, or X-rays may be cancelled for obese patients because the equipment is not compatible. Here the results of analysis of the

data with these participants may be different from the results if they are excluded; however, a significant change to the direction of the effect is not expected.[19]

Nevertheless, the most common pattern of missing data is directly related to the behavior of the variable being studied. For example, patients suffering little pain are more likely to conclude a questionnaire on symptoms; dropping out of a study might be more common among those who experience adverse effects or in a placebo group (less clinical effect); or even, more severe hypertensive patients may not attend visits to have blood pressure measured, because they are more likely to have to attend the emergency room or because of headaches. These data are missing not at random (MNAR), and they cause serious selection bias in a sample, compromising generalization of results.

If there is a small percentage of missing data and they have a random pattern (MAR or MCAR), there are a number of options for imputation. Data with a non-random pattern of absence (MNAR) demand for support from a statistical professional with experience in identification and treatment of these data.

Exclusion of the full record (all data) for participants that have missing data values (casewise or listwise) reduces the total sample size and can penalize the inferential power of the analysis if the sample is small, or, in cases in which the pattern of absence is non-random (MNAR), it can cause analytical bias. One option is to only exclude the subject from analyses of the missing variables (pairwise), reducing the sample size of descriptive statistics for these variables only or in analyses (for example, correlations) that employ that variable, allowing the remainder of the data available on the subject to be used in other statistical analyses.[22]

Substitution of the missing value by an estimator of the central tendency (mean, mode, or median) of the other values for the variable is a relatively precise option, but it reduces the variability of data (overfit) and does not consider the effect of other variables in imputation. On the other hand, substitution of the missing value by the value in the adjacent record (value for the previous or next subject) increases the variability of data (underfit), and also does not take other variables into account. Use of multivariate regression techniques to estimate the missing value as a function of the remaining variables offers the most precise estimation, but reduces the variability of the data (overfit). These options are most appropriate when the magnitude of missing data is small (< 5%).

The best technique for substitution of absent values is multiple imputation, which employs several predictive models to validate values by testing a selection of different missing data, in order to maintain the same variance as the available values for the variable (minimizing overfit). Multiple imputation of absent values gives better analytical performance than exclusion of cases (listwise) or variables (pairwise) with missing values. In general, the multiple imputation model should contain all of the study variables, and at least 10 attempts (iterations) should be run to arrive at the best estimation of the missing data.[23-29]

Returning to the example in Table 1, the correlation between values for systolic blood pressure and body mass index is $\rho = 0.60$ ($p = 0.01$) for the 17 original pairs of data, and $\rho = 0.61$ ($p < 0.01$) after multiple imputation of the two missing values.[30] These values show that multiple imputation techniques do not interfere with the magnitude of the effect (for example, Spearman's ρ, odds ratios, β coefficients of regressions), but they do increase the analytical power and the precision of estimates.[21,27]

It is important to point out that these multiple imputation are not applicable to studies of just one variable, losses with a MNAR pattern, or for when the intention is to increase (artificially) the sample size. Additionally, imputation of the dependent variable (principal study outcome) on the basis of its covariates is not recommended.[29,31]

There is a special case of missing data which is the set of data that is lost because of participants who leave the study. These events are known as dropouts and they are the cause of a profusion of academic discussions on analysis of longitudinal studies (for example, cohorts and clinical trials).[32-39] Nevertheless, as mentioned earlier, dropouts or losses to follow-up exceeding 10% of participants can seriously compromise the results of a study, except in survival studies, in which the principal outcome is itself time of survival.[40] Dropouts can also be the result of events, which may or may not be linked to other study variables (for example, failure to attend because of an adverse event related to treatment), and analysis of the results of a study with exclusion of participants that drop out (per protocol analysis) can give a false estimate of the effect or safety of a treatment.[34,35,41]

Longitudinal intervention studies (for example, randomized clinical trials) should preferably analyze all participants by intention to treat (ITT), so that all of those randomized and allocated to a group should be analyzed at the end of the study, irrespective of diversions from the therapeutic protocol (for example, withdrawal or change of treatment) or of dropouts. For dropout cases, one option for ITT analysis of missing dependent variables is to copy the value from the subject's last visit, known as last observed carried forward (LOCF), although it tends to underfit

estimations of the parameter and can reduce the effect of treatment.[42,43] Recovering the information is preferable to LOCF, even on a date long after that scheduled for the visit. Additionally, some techniques for analysis of longitudinal studies (generalized linear mixed-effects models) can deal with missing data and dropouts in their analytical structures.[35,37,39,44-48]

In general, descriptive statistics and bivariate analyses should be conducted including the outlier values (untransformed) and should also consider missing data, to preserve the fidelity of the description of the original sample. The techniques described here are preferred to ensure successful multivariate analyses, where the existence of outlier values or missing data can violate the preconditions of the statistical tests (for example, normality) or require exclusion of subjects and variables from the study.

Finally, the strategies used to deal with missing data and outliers should be described in detail in the methodology and when presenting the results. Irrespective, it is a good practice to conduct an analysis of the sensitivity of the results, running the same data analyses with the original values and after exclusion of cases with missing data and outliers, to test whether the direction of the results is aligned with the conclusions reached at using corrected data.[21,36,49,50]

REFERENCES

1. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol. 2017;70(4):407-11. http://dx.doi.org/10.4097/kjae.2017.70.4.407. PMid:28794835.
2. Norman GR, Streiner DL. Biostatistics. The bare essentials. 4th ed. Shelton: People's Medical Publishing House; 2014.
3. Miot HA. Agreement analysis in clinical and experimental trials. J Vasc Bras. 2016;15:89-92. http://dx.doi.org/10.1590/1677-5449.004216. PMid:29930571.
4. Miot HA. Assessing normality of data in clinical and experimental trials. J Vasc Bras. 2017;16:88-91. http://dx.doi.org/10.1590/1677-5449.041117. PMid:29930631.
5. de Cheveigné A, Arzounian D. Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. Neuroimage. 2018;172:903-12. http://dx.doi.org/10.1016/j.neuroimage.2018.01.035. PMid:29448077.
6. Penny KI, Jolliffe IT. Multivariate outlier detection applied to multiply imputed laboratory data. Stat Med. 1999;18:1879-95.
7. Ramsay T, Elkum N. A comparison of four different methods for outlier detection in bioequivalence studies. J Biopharm Stat. 2005;15(1):43-52. http://dx.doi.org/10.1081/BIP-200040815. PMid:15702604.
8. Abellana Sangra R, Farran Codina A. The identification, impact and management of missing values and outlier data in nutritional epidemiology. Nutr Hosp. 2015;31(Suppl 3):189-95. PMid:25719786.
9. Shete S, Beasley TM, Etzel CJ, et al. Effect of winsorization on power and type 1 error of variance components and related methods of QTL detection. Behav Genet. 2004;34(2):153-9. http://dx.doi.org/10.1023/B:BEGE.0000013729.26354.da. PMid:14755180.
10. Ramalle-Gomara E, Andres De Llano JM. Use of robust methods in inferential statistics. Aten Primaria. 2003;32(3):177-82. PMid:12975106.
11. Evans K, Love T, Thurston SW. Outlier identification in model-based cluster analysis. J Classif. 2015;32(1):63-84. http://dx.doi.org/10.1007/s00357-015-9171-5. PMid:26806993.
12. Wilcox RR. Robust ANCOVA using a smoother with bootstrap bagging. Br J Math Stat Psychol. 2009;62(Pt 2):427-37. http://dx.doi.org/10.1348/000711008X325300. PMid:18652737.
13. O'Hagan A, Stevens JW. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? Health Econ. 2003;12(1):33-49. http://dx.doi.org/10.1002/hec.699. PMid:12483759.
14. Jiang X, Guo X, Zhang N, Wang B, Zhang B. Robust multivariate nonparametric tests for detection of two-sample location shift in clinical trials. PLoS One. 2018;13(4):e0195894. http://dx.doi.org/10.1371/journal.pone.0195894. PMid:29672555.
15. Cleophas TJ. Clinical trials: robust tests are wonderful for imperfect data. Am J Ther. 2015;22(1):e1-5. http://dx.doi.org/10.1097/MJT.0b013e31824c3ee1. PMid:23896742.
16. Wagstaff DA, Elek E, Kulis S, Marsiglia F. Using a nonparametric bootstrap to obtain a confidence interval for Pearson's r with cluster randomized data: a case study. J Prim Prev. 2009;30(5):497-512. http://dx.doi.org/10.1007/s10935-009-0191-y. PMid:19685290.
17. Rascati KL, Smith MJ, Neilands T. Dealing with skewed data: an example using asthma-related costs of medicaid clients. Clin Ther. 2001;23(3):481-98. http://dx.doi.org/10.1016/S0149-2918(01)80052-7. PMid:11318082.
18. Vickers AJ, Altman DG. Statistics notes: missing outcomes in randomised trials. BMJ. 2013;346:1-4. http://dx.doi.org/10.1136/bmj.f3438. PMid:23744649.
19. Altman DG, Bland JM. Missing data. BMJ. 2007;334(7590):424. http://dx.doi.org/10.1136/bmj.38977.682025.2C. PMid:17322261.
20. Miot HA, Medeiros LM, Siqueira CRS, et al. Association between coronary artery disease and the diagonal earlobe and preauricular creases in men. An Bras Dermatol. 2006;81:29-33. http://dx.doi.org/10.1590/S0365-05962006000100003.
21. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. http://dx.doi.org/10.1136/bmj.b2393. PMid:19564179.
22. Little RJ. Regression with missing X's: A review. J Am Stat Assoc. 1992;87:1227-37.
23. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157-66. http://dx.doi.org/10.2147/CLEP.S129785. PMid:28352203.
24. Enders CK. Multiple imputation as a flexible tool for missing data handling in clinical research. Behav Res Ther. 2017;98:4-18. http://dx.doi.org/10.1016/j.brat.2016.11.008. PMid:27890222.
25. Stanimirova I, Walczak B. Classification of data with missing elements and outliers. Talanta. 2008;76(3):602-9. http://dx.doi.org/10.1016/j.talanta.2008.03.049. PMid:18585327.
26. Mackinnon A. The use and reporting of multiple imputation in medical research - a review. J Intern Med. 2010;268(6):586-93. http://dx.doi.org/10.1111/j.1365-2796.2010.02274.x. PMid:20831627.
27. Harel O, Mitchell EM, Perkins NJ, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):576-84. http://dx.doi.org/10.1093/aje/kwx349. PMid:29165547.
28. Enders CK. Multiple imputation as a flexible tool for missing data handling in clinical research. Behav Res Ther. 2017;98:4-18. http://dx.doi.org/10.1016/j.brat.2016.11.008. PMid:27890222.

29. Nunes LN, Klück MM, Fachel JMG. Multiple imputations for missing data: a simulation with epidemiological data. Cad Saude Publica. 2009;25(2):268-78. http://dx.doi.org/10.1590/S0102-311X2009000200005. PMid:19219234.

30. Miot HA. Correlation analysis in clinical and experimental studies. J Vasc Bras. 2018;17(4):275-9. http://dx.doi.org/10.1590/1677-5449.174118. PMid:30787944.

31. Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat Methods Med Res. 2018;27(9):2610-26. http://dx.doi.org/10.1177/0962280216683570. PMid:28034175.

32. Gades NM, Jacobson DJ, McGree ME, et al. Dropout in a longitudinal, cohort study of urologic disease in community men. BMC Med Res Methodol. 2006;6(1):58. http://dx.doi.org/10.1186/1471-2288-6-58. PMid:17169156.

33. Curran D, Molenberghs G, Aaronson NK, Fossa SD, Sylvester RJ. Analysing longitudinal continuous quality of life data with dropout. Stat Methods Med Res. 2002;11(1):5-23. http://dx.doi.org/10.1191/0962280202sm270ra. PMid:11923994.

34. Cheng J, Edwards LJ, Maldonado-Molina MM, Komro KA, Muller KE. Real longitudinal data analysis for real people: building a good enough mixed model. Stat Med. 2010;29(4):504-20. PMid:20013937.

35. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med. 2013;86(3):343-58. PMid:24058309.

36. Moreno-Betancur M, Chavance M. Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. Stat Methods Med Res. 2016;25(4):1471-89. http://dx.doi.org/10.1177/0962280213490014. PMid:23698867.

37. Rombach I, Jenkinson C, Gray AM, Murray DW, Rivero-Arias O. Comparison of statistical approaches for analyzing incomplete longitudinal patient-reported outcome data in randomized controlled trials. Patient Relat Outcome Meas. 2018;9:197-209. http://dx.doi.org/10.2147/PROM.S147790. PMid:29950913.

38. Garcia TP, Marder K. Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model. Curr Neurol Neurosci Rep. 2017;17(2):14. http://dx.doi.org/10.1007/s11910-017-0723-4. PMid:28229396.

39. Edwards LJ. Modern statistical techniques for the analysis of longitudinal data in biomedical research. Pediatr Pulmonol. 2000;30(4):330-44. http://dx.doi.org/10.1002/1099-0496(200010)30:4<330::AID-PPUL10>3.0.CO;2-D. PMid:11015135.

40. Miot HA. Survival analysis in clinical and experimental studies. J Vasc Bras. 2017;16:267-9. http://dx.doi.org/10.1590/1677-5449.001604. PMid:29930659.

41. Little R, Kang S. Intention-to-treat analysis with treatment discontinuation and missing data in clinical trials. Stat Med. 2015;34(16):2381-90. http://dx.doi.org/10.1002/sim.6352. PMid:25363683.

42. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ. 2011;342:1-9. http://dx.doi.org/10.1136/bmj.d40. PMid:21300711.

43. Streiner D, Geddes J. Intention to treat analysis in clinical trials when there are missing data. Evid Based Ment Health. 2001;4(3):70-1. http://dx.doi.org/10.1136/ebmh.4.3.70. PMid:12004740.

44. Bagatin E, Miot HA. How to design and write a clinical research protocol in Cosmetic Dermatology. An Bras Dermatol. 2013;88(1):69-75. http://dx.doi.org/10.1590/S0365-05962013000100008. PMid:23539006.

45. Resseguier N, Giorgi R, Paoletti X. Sensitivity analysis when data are missing not-at-random. Epidemiology. 2011;22(2):282. http://dx.doi.org/10.1097/EDE.0b013e318209dec7. PMid:21293212.

46. Yamaguchi Y, Misumi T, Maruo K. A comparison of multiple imputation methods for incomplete longitudinal binary data. J Biopharm Stat. 2018;28(4):645-67. http://dx.doi.org/10.1080/10543406.2017.1372772. PMid:28886277.

47. Wen L, Terrera GM, Seaman SR. Methods for handling longitudinal outcome processes truncated by dropout and death. Biostatistics. 2018;19(4):407-25. http://dx.doi.org/10.1093/biostatistics/kxx045. PMid:29028922.

48. Spratt M, Carpenter J, Sterne JA, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. 2010;172(4):478-87. http://dx.doi.org/10.1093/aje/kwq137. PMid:20616200.

49. Ferretti F, Saltelli A, Tarantola S. Trends in sensitivity analysis practice in the last decade. Sci Total Environ. 2016;568:666-70. http://dx.doi.org/10.1016/j.scitotenv.2016.02.133. PMid:26934843.

50. Tseng CH, Elashoff R, Li N, Li G. Longitudinal data analysis with non-ignorable missing data. Stat Methods Med Res. 2016;25(1):205-20. http://dx.doi.org/10.1177/0962280212448721. PMid:22637472.

Correspondence
Hélio Amante Miot
Universidade Estadual Paulista – UNESP
Av. Prof. Mário Rubens Guimarães Montenegro, s/n - Distrito de Rubião Junior
CEP 18618-687 - Botucatu (SP), Brasil
Tel.: +55 (14) 3882-4922
E-mail: heliomiot@gmail.com

Author information
HAM - Tenured professor, Departamento de Dermatologia e Radioterapia, Faculdade de Medicina de Botucatu, Universidade Estadual Paulista (UNESP), Botucatu, SP, Brasil.

ARTIGO DE REVISÃO

JORNAL **VASCULAR** BRASILEIRO
ISSN 1677-7301 (Online)

Valores anômalos e dados faltantes em estudos clínicos e experimentais

Anomalous values and missing data in clinical and experimental studies

Hélio Amante Miot[1]

Resumo

Durante a análise dos dados de uma pesquisa científica, é habitual deparar-se com valores anômalos ou dados faltantes. Valores anômalos podem ser resultado de erros de registro, de digitação, de aferição instrumental, ou configurarem verdadeiros *outliers*. Nesta revisão, são discutidos conceitos, exemplos e formas de identificar e de lidar com tais contingências. No caso de dados faltantes, discutem-se técnicas de imputação dos valores para evitar a exclusão do sujeito da pesquisa, caso não seja possível recuperar a informação das fichas de registro ou reabordar o participante.

Palavras-chave: análise de dados; base de dados; discrepância; imputação múltipla.

Abstract

During analysis of scientific research data, it is customary to encounter anomalous values or missing data. Anomalous values can be the result of errors of recording, typing, measurement by instruments, or may be true outliers. This review discusses concepts, examples and methods for identifying and dealing with such contingencies. In the case of missing data, techniques for imputation of the values are discussed in, order to avoid exclusion of the research subject, if it is not possible to retrieve information from registration forms or to re-address the participant.

Keywords: data analysis; database; outlier; multiple imputation.

Como citar: Miot HA. Valores anômalos e dados faltantes em estudos clínicos e experimentais. J Vasc Bras. 2019;18: e20190004. https://doi.org/10.1590/1677-5449.190004

[1]Universidade Estadual Paulista – UNESP, Faculdade de Medicina de Botucatu, Departamento de Dermatologia e Radioterapia, Botucatu, SP, Brasil.
Fonte de financiamento: Nenhuma.
Conflito de interesse: Os autores declararam não haver conflitos de interesse que precisam ser informados.
Submetido em: Janeiro 08, 2019. Aceito em: Março 14, 2019.

O estudo foi realizado na Faculdade de Medicina de Botucatu, Universidade Estadual Paulista (UNESP), Botucatu, SP, Brasil.

Antes de iniciar o processo de análise dos dados de uma pesquisa clínica ou biomédica, é imperioso avaliar cuidadosamente a existência de dados faltantes ou valores anômalos na amostra, pois eles são habituais, e a sua inobservância pode comprometer as conclusões do estudo ou seu poder inferencial[1]. Valores anômalos podem ser resultado de erros de registro, de digitação, de aferição instrumental ou configurarem verdadeiros *outliers*[2].

À medida que a amostra e/ou o número de variáveis aumenta, cresce a chance de ocorrerem erros de digitação. Em estudos com amostras vultosas, utilizam-se inclusive artifícios como dupla digitação ou revisão amostral dos registros, para identificar (e prevenir) possíveis erros.

A Tabela 1 apresenta dados hipotéticos de uma pesquisa clínica em que ocorrem alguns padrões de valores anômalos, *outliers* e dados faltantes.

A observação da sequência dos identificadores evidencia que o participante número 8 não foi incluído nos registros da Tabela 1, o que pode decorrer da exclusão protocolar ou de falha do digitador.

Na coluna da idade, há um participante com 555 anos, uma provável digitação múltipla de um algarismo (por exemplo, 55 *versus* 555 anos). Todavia, caso ocorresse troca de algarismos que resultasse em um valor coerente (como 23 *versus* 32 anos, ou 4 *versus* 44 anos), a identificação visual do erro seria muito dificultada.

Ainda sobre o registro da idade, há tendência dos sujeitos da pesquisa referirem sua idade arredondada para um valor abaixo do real (por exemplo, 40 em vez de 43 anos). A fim de minimizar esse tipo de viés, recomenda-se o registro do ano de nascimento, ou mesmo a data completa, sendo a idade calculada posteriormente na fase de análise dos dados. Nesse caso, deve-se ter atenção de não registrar a data ou ano atuais em vez da data ou ano de nascimento do participante (por exemplo, 2019 em vez de 1979).

O sexo do participante 17 foi registrado como "N", uma codificação não utilizada para essa variável (M ou F). Visto que "N" é uma letra adjacente ao "M" no teclado, trata-se também de um padrão habitual de falha na digitação. Além disso, os sistemas usados para análise estatística podem diferenciar maiúsculas de minúsculas (por exemplo, "M" de "m") e também a acentuação (por exemplo, "não" e "nao"). Tais contingências são prevenidas pela codificação numérica das respostas (por exemplo, Masculino = 1 e Feminino = 2; Sim = 1 e Não = 2).

Eventualmente, a percepção do erro depende da avaliação de mais variáveis, como ocorre no registro 16, em que um participante do sexo masculino refere três gestações. Da mesma forma, no registro 7, uma participante de apenas 18 anos refere seis gestações. Por fim, o participante 21 apresenta exatamente os mesmos registros que o participante 9 em todas as variáveis, sugerindo dupla inclusão no estudo.

Tabela 1. Exemplo de registro de dados (hipotéticos) de uma pesquisa clínica.

Identificador	Idade*	Sexo	Gestações	Pressão arterial sistólica**	Pressão arterial diastólica**	Índice de massa corporal#
1	46	F	2	120	80	23,8
2	50	F	3		110	24,9
3	69	F		110	150	22,9
4	22	M	0	135	85	24,1
5	555	M	0	165	95	27,0
6	38		0	125	75	23,9
7	18	F	6	155	90	26,1
9	58	F	3	135	75	24,2
10		M	0	145	85	25,8
11	93	M	0	150	115	24,1
12	45	F	1	135	135	23,7
13	43	F	1	120	80	25,1
14	38		2	140	90	25,0
15	37	M	0	235	180	29,2
16	30	M	3	130	100	24,9
17	42	N	0	120	70	23,8
18	30	F		115	75	
19	25	F	0	135	100	24,2
20	28	M	0	145	105	23,1
21	58	F	3	135	75	24,2

* Idade em anos completos; ** Pressão arterial em mmHg; # Índice de massa corporal em kg/m^2.

Incongruências também devem ser verificadas diante de valores que apresentem comportamentos dependentes. No caso, a pressão arterial diastólica deve ser menor que a sistólica, o que não se observa nos registros 3 e 12, nos quais ocorreu inversão dos registros e duplicação do valor digitado, respectivamente.

Erros de aferição de diferentes instrumentos (por exemplo, esfigmomanômetros e balanças) induzem ao erro sistemático, muito difícil de ser identificado e corrigido. Quando tal erro se propaga homogeneamente na amostra (por exemplo, aferição de 10 mmHg a mais para todos os registros), não acarreta grande prejuízo na comparação interna dos grupos. Entretanto, quando se usam diferentes instrumentos com problemas de calibragem ou baixa reprodutibilidade, há aumento na variabilidade e perda da exatidão dos parâmetros. É fundamental a preocupação com a concordância dos instrumentos de coleta dos dados ou dos métodos laboratoriais, pois a correção desses vieses na fase analítica (por exemplo, transformação em valores de Z-score para os dados de cada instrumento) tem performance insatisfatória[3].

Nesse ínterim, é conveniente comentar que algumas informações imbuídas de algum valor cultural podem ser falsamente reportadas pelos participantes, com a finalidade de aceitação, identificação social ou de julgamento moral. De uma maneira geral, peso corporal, uso de substâncias ilícitas e número de relacionamentos extraconjugais tendem a ser subestimados pelos sujeitos da pesquisa, enquanto a estatura, prática de sexo seguro e atitudes afirmativas (por exemplo, altruísmo, solidariedade ou bom senso) são reportados acima dos reais valores. Não há uma forma infalível de prevenir esse tipo de falso relato, tampouco a estatística pode corrigir tais vieses. Entretanto, os pesquisadores recomendam, além de medidas objetivas (por exemplo, medir o peso e a altura durante a entrevista, verificar o ano de nascimento em documento ou registro hospitalar), o uso de questões/perguntas confirmatórias que permitam verificar a integridade da informação (por exemplo, no início da pesquisa questionar o número de vezes por mês que usa substância ilícita e, ao final da entrevista, questionar o número de vezes por semana que usa substâncias, discriminando maconha, cocaína, ácido, etc.).

A exatidão dos registros é primordial para a qualidade do estudo e a validade das suas conclusões; dessa forma, o planejamento da pesquisa deve considerar esforços para minimizar esse tipo de situação.

Existem, ainda, valores que se distanciam muito do comportamento da amostra, os chamados *outliers*, que não se referem a erros de registro mas sim ao encontro probabilístico de valores extremos (para mais ou para menos) que realmente existem na população. No exemplo da Tabela 1, o participante 11 tem 93 anos, e o participante 15 apresenta níveis pressóricos discrepantes dos demais.

Em distribuições normais[4], valores *outliers* são definidos como os mais extremos que 1,5 desvio interquartílico abaixo do p25 ou acima do p75 de uma amostra (Figura 1), ou valores padronizados que ultrapassem três desvios padrão (para mais ou para menos) na amostra. Em distribuições não normais, análises de correlações ou análises multivariadas, a identificação de valores *outliers* é mais complexa, e ultrapassa o escopo desta revisão[5-7].

Além da identificação de *outliers*, há grande discussão sobre como lidar com esses dados. Se, por um lado,

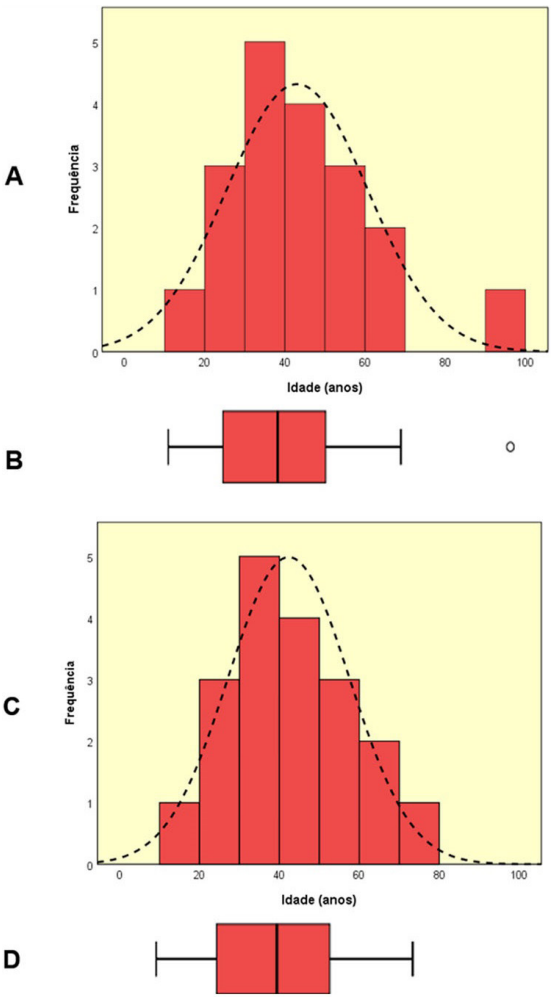

Figura 1. Histograma e diagrama box plot da variável idade apresentada na Tabela 1 (A e B) e após sua winsorização (C e D). Houve um valor outlier – a idade de 93 anos –, situado mais extremamente que 1,5 vezes o desvio interquartílico (25 anos) somado ao percentil 75 (55 anos), que foi winsorizado para 70 anos (n = 19).

esses registros destoam da amostra, aumentam a variabilidade dos dados, comprometem a normalidade da distribuição, reduzem o poder estatístico e influenciam na inferência populacional, por outro lado, são valores reais, de sujeitos que participam da população do estudo. *Outliers* podem, inclusive, ser indicadores de padrões especiais dentro da amostra, fundamentando novas hipóteses para o fenômeno estudado, ou ainda revelar distribuições de probabilidade não normais subjacentes da população[8].

Os testes estatísticos bivariados habituais para dados paramétricos (teste *t* de Student, ANOVA, coeficiente de correlação de Pearson) são relativamente robustos para lidar com uma pequena frequência de *outliers*. Já os testes baseados em postos (Mann-Whitney, Wilcoxon, Kruskal-Wallis, coeficiente de Spearman) não são afetados por valores extremos. Dessa forma, a decisão de excluir sujeitos com valores *outliers* penaliza a amostra, e deve ser evitada. Em vez disso, se necessário, é possível tratar *outliers* com winsorização ou *trimming*, ou empregar técnicas de agrupamento (clusterização), reamostragem (*bootstrap*) ou análises estatísticas robustas, que promovem uma aproximação para uma distribuição de probabilidade, baseada nos dados centrais[9-13].

Na winsorização, o dado aberrante é substituído por um valor que supere o seu antecessor, tornando o *outlier* mais próximo do conjunto de dados[1]. No caso da Tabela 1, a idade de 93 anos poderia ser winsorizada para 70 anos, uma unidade acima do penúltimo valor mais alto: 69 anos (Figura 1).

No *trimming*, uma porcentagem dos extremos da amostra (por exemplo, os 2% mais extremos) é excluída bilateralmente da análise. Esse procedimento homogeneíza a amostra, mas pode penalizar o poder da análise estatística, por reduzir o tamanho amostral[1].

As técnicas de agrupamento avaliam padrões de proximidade dos participantes baseados no comportamento das demais variáveis, e o valor *outlier* é substituído pela média verificada entre os sujeitos identificados como um grupo. Tanto as técnicas de clusterização e imputação baseada em reamostragem quanto os métodos estatísticos robustos exigem o envolvimento de um profissional estatístico experiente[10,11,14-17].

É importante que os pesquisadores adotem rotinas de identificação de valores anômalos e *outliers*, devido ao ônus inferencial que eles infligem, especialmente em estudos com pequeno número de participantes. Caso *outliers* ocorram em baixa frequência na amostra e não modifiquem a conclusão da análise, recomenda-se não promover nenhuma transformação dos dados.

Outra situação habitual à pesquisa clínica e experimental são dos dados faltantes, que são facilmente diagnosticados visualmente, pelo "vazio" que infligem à planilha de dados (Tabela 1). Contudo, à medida que o número de sujeitos e/ou de variáveis aumenta, recomenda-se também utilizar estratégias de verificação da sua ocorrência. Ademais, algumas planilhas e softwares de análise substituem automaticamente os dados faltantes por ZERO ou por um valor aberrante (por exemplo, 999), o que pode acarretar ainda mais prejuízo caso esses dados não sejam identificados.

Os dados faltantes podem se originar tanto de erros de digitação quanto da sua real indisponibilidade durante a coleta. Caso se possa recuperar as fichas de registro ou reabordar o sujeito da pesquisa, essas são as melhores alternativas para essa contingência. Em alguns casos, o comportamento de outras variáveis permite deduzir com certeza o dado faltante. Na Tabela 1, o registro 14 deve se tratar de uma mulher, já que refere duas gestações[2].

Entretanto, alguns dados não podem ser recuperados *a posteriori* (por exemplo, paciente faleceu, camundongos foram sacrificados), não podem ser deduzidos, sofrem mudança com o momento da coleta, ou resultam de experimentos mais complexos. Essas circunstâncias demandam o uso de certas técnicas estatísticas para lidar com essas limitações[18-20].

O primeiro passo para tratar os dados faltantes é a análise da magnitude da ausência dos valores. Sujeitos com mais de 10% dos dados faltantes, ou variáveis com mais de 10% dos valores faltantes, são situações desfavoráveis ao uso de técnicas de imputação de valores, e a permanência desse sujeito ou dessa variável no estudo deve ser questionada.

O segundo passo é a avaliação do padrão de dados faltantes, já que as técnicas de imputação exigem que ausência de dados tenha uma certa independência frente às variáveis subjacentes, pois a própria falta da informação pode estar ligada ao comportamento de alguma variável.

Dados faltantes que não seguem um padrão de ausência são chamados de valores faltantes completamente aleatórios (*missing completely at random*, MCAR), como quando uma folha do questionário é perdida, uma amostra de sangue coagula, ou um paciente se muda de cidade. Nesse caso, assume-se que as ausências de respostas decorram de elementos externos ao protocolo, e que a análise dos dados com ou sem os participantes com dados faltantes não muda a dimensão do efeito[21].

Existem também os dados faltantes aleatórios (*missing at random*, MAR), em que a falta de um valor decorre do efeito de outra covariável secundária: os menos instruídos podem deixar itens sem resposta por baixa compreensão, questões de ordem sexual podem ser negligenciadas por participantes promíscuos, ou exames radiológicos podem ser cancelados em

pacientes obesos por incompatibilidade do aparelho. Aqui, a análise dos dados com os participantes pode ser algo divergente dos resultados após a exclusão desses sujeitos; entretanto, não se espera uma modificação importante na direção do efeito[19].

Contudo, o padrão mais habitual de dados faltantes é diretamente relacionado ao próprio comportamento da variável estudada. Por exemplo, a conclusão de um questionário de sintomas é mais provável em pacientes com pouca dor; o abandono do estudo é mais comum em quem tem mais efeitos adversos ou no grupo placebo (menor efeito clínico); ou ainda, as visitas para aferição de pressão arterial podem ser perdidas entre os hipertensos mais graves, devido à maior necessidade de visitas ao pronto-socorro ou à ocorrência de cefaleia. Trata-se da perda não aleatória (*missing not -at random*, MNAR), e inflige importante viés de seleção na amostra, comprometendo a generalização dos resultados.

Caso haja uma pequena porcentagem de dados faltantes e eles apresentem um padrão aleatório (MAR ou MCAR), há diferentes opções de imputação descritas. Dados com padrão de ausência não aleatório (MNAR) demandam suporte de um profissional estatístico experiente na identificação e no tratamento dos dados.

A exclusão do registro completo (todos os dados) do participante que possua algum dado faltante (*casewise* ou *listwise*) reduz a amostra total e pode penitenciar o poder inferencial da análise quando a amostra for pequena, ou, em casos que o padrão de ausência seja não aleatório (MNAR), pode incluir viés analítico. Há, contudo, a opção de se excluir o sujeito da análise exclusiva das variáveis ausentes (*pairwise*), reduzindo o tamanho amostral apenas na estatística descritiva dessas variáveis ou em análises (por exemplo, correlações) que utilizem aquela variável, o que leva ao aproveitamento dos demais dados completos do sujeito para as demais técnicas estatísticas[22].

A substituição do valor faltante por um estimador de tendência central (média, moda ou mediana) dos demais valores da variável é uma alternativa relativamente precisa, mas reduz a variabilidade dos dados (*overfit*) e não considera o efeito das demais variáveis na imputação. Por outro lado, a substituição do valor faltante pelo dado registrado adjacente (valor do sujeito anterior ou posterior) promove aumento da variabilidade dos dados (*underfit*), e também não pondera as demais variáveis. O uso de alguma técnica de regressão múltipla para estimar o valor faltante em função das demais variáveis apresenta melhor precisão da estimativa, mas reduz a variabilidade dos dados (*overfit*). Essas alternativas são mais indicadas quando a ausência de dados é de pequena magnitude (< 5%).

A técnica mais indicada na substituição de valores ausentes chama-se imputação múltipla, que utiliza diferentes modelos preditivos para validar os valores a partir da testagem de diferentes dados faltantes, a fim de manter a mesma variância dos valores na variável (minimiza o *overfit*). Imputação múltipla de valores ausentes resulta em melhor performance analítica que a exclusão dos casos (*listwise*) ou das variáveis faltantes (*pairwise*). Em geral, deve-se incluir no modelo de imputação múltipla todas as variáveis do estudo, e devem ser realizadas até 10 tentativas (iterações) para a melhor estimativa dos dados faltantes[23-29].

No exemplo da Tabela 1, a correlação entre os valores de pressão arterial sistólica e índice de massa corporal é $\rho = 0,60$ ($p = 0,01$) para os 17 pares de dados originais, e $\rho = 0,61$ ($p < 0,01$) após imputação múltipla dos dois valores faltantes[30]. Esses valores evidenciam que técnicas de imputação múltipla não interferem com a dimensão do efeito (por exemplo, ρ de Spearman, *odds ratio*, coeficiente β da regressão), apenas aumentam o poder da análise e a precisão das estimativas[21,27].

É importante salientar que essas técnicas de imputação múltipla não se aplicam a estudos com apenas uma variável, perdas com padrão MNAR, ou quando se tem intenção de ampliar (artificialmente) o tamanho de amostra. Da mesma forma, a imputação da variável dependente (desfecho principal do estudo) em função das demais covariáveis não é recomendada[29,31].

Uma situação especial se refere ao conjunto de dados faltantes decorrentes de abandono de seguimento do estudo. Esses eventos são chamados *dropouts*, e dão origem a uma rica discussão acadêmica sobre a análise de estudos longitudinais (por exemplo, coorte e ensaios clínicos)[32-39]. Da mesma forma, como discutido anteriormente, abandonos ou perdas no seguimento de mais de 10% dos participantes podem comprometer seriamente os resultados do estudo, exceto em ensaios de sobrevivência, em que o desfecho principal é, propriamente, o tempo de sobrevida[40]. *Dropouts* também podem ocorrer devido a eventos ligados ou não às demais variáveis do estudo (por exemplo, falta na visita do estudo por um evento adverso do tratamento), e a análise dos resultados de um estudo com a exclusão dos participantes *dropouts* (*per protocol analysis*) pode promover uma falsa estimativa do efeito ou da segurança do tratamento[34,35,41].

Estudos longitudinais de intervenção (por exemplo, ensaios clínicos randomizados) devem, preferencialmente, ter seus participantes analisados por intenção de tratamento (*intention to treat*, ITT), em que todos os randomizados e alocados em um grupo devem ser analisados ao final do estudo, independentemente de desvios do protocolo terapêutico (por exemplo,

descontinuidade ou troca do tratamento) ou de *dropouts*. Em casos de *dropouts*, uma alternativa para a análise ITT de variáveis dependentes faltantes é replicar o valor da última visita do sujeito (*last observed carried forward,* LOCF), o que promove *underfit* na estimativa do parâmetro e pode reduzir o efeito do tratamento[42,43]. A recuperação da informação, mesmo distante da data prevista da visita, é preferível ao LOCF. Além disso, algumas técnicas de análise de estudos longitudinais (modelos lineares generalizados de efeitos mistos) lidam com dados faltantes e *dropouts* nas suas estruturas analíticas[35,37,39,44-48].

De forma geral, a estatística descritiva e as análises bivariadas devem ser conduzidas com os valores *outliers* (não transformados) e considerar os dados faltantes, para se manter a fidedignidade com a descrição da amostra original. As técnicas aqui descritas são preferenciais para o sucesso das análises multivariadas, em que a existência de valores *outliers* ou dados faltantes compromete os pré-requisitos dos testes estatísticos (por exemplo, normalidade) ou implicam na exclusão de sujeitos e de variáveis da pesquisa.

Finalmente, as estratégias de tratamento para dados faltantes e *outliers* também devem ser detalhadamente descritas na metodologia e na apresentação dos resultados. Ademais, é recomendável promover uma análise de sensibilidade dos resultados, procedendo a mesma análise dos dados considerando os valores originais e excluindo os casos com dados faltantes e *outliers*, a fim de identificar se a direção dos resultados segue a mesma das conclusões com os dados corrigidos[21,36,49,50].

REFERÊNCIAS

1. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol. 2017;70(4):407-11. http://dx.doi.org/10.4097/kjae.2017.70.4.407. PMid:28794835.
2. Norman GR, Streiner DL. Biostatistics. The bare essentials. 4th ed. Shelton: People's Medical Publishing House; 2014.
3. Miot HA. Agreement analysis in clinical and experimental trials. J Vasc Bras. 2016;15:89-92. http://dx.doi.org/10.1590/1677-5449.004216. PMid:29930571.
4. Miot HA. Assessing normality of data in clinical and experimental trials. J Vasc Bras. 2017;16:88-91. http://dx.doi.org/10.1590/1677-5449.041117. PMid:29930631.
5. de Cheveigné A, Arzounian D. Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. Neuroimage. 2018;172:903-12. http://dx.doi.org/10.1016/j.neuroimage.2018.01.035. PMid:29448077.
6. Penny KI, Jolliffe IT. Multivariate outlier detection applied to multiply imputed laboratory data. Stat Med. 1999;18:1879-95.
7. Ramsay T, Elkum N. A comparison of four different methods for outlier detection in bioequivalence studies. J Biopharm Stat. 2005;15(1):43-52. http://dx.doi.org/10.1081/BIP-200040815. PMid:15702604.
8. Abellana Sangra R, Farran Codina A. The identification, impact and management of missing values and outlier data in nutritional epidemiology. Nutr Hosp. 2015;31(Suppl 3):189-95. PMid:25719786.
9. Shete S, Beasley TM, Etzel CJ, et al. Effect of winsorization on power and type 1 error of variance components and related methods of QTL detection. Behav Genet. 2004;34(2):153-9. http://dx.doi.org/10.1023/B:BEGE.0000013729.26354.da. PMid:14755180.
10. Ramalle-Gomara E, Andres De Llano JM. Use of robust methods in inferential statistics. Aten Primaria. 2003;32(3):177-82. PMid:12975106.
11. Evans K, Love T, Thurston SW. Outlier identification in model-based cluster analysis. J Classif. 2015;32(1):63-84. http://dx.doi.org/10.1007/s00357-015-9171-5. PMid:26806993.
12. Wilcox RR. Robust ANCOVA using a smoother with bootstrap bagging. Br J Math Stat Psychol. 2009;62(Pt 2):427-37. http://dx.doi.org/10.1348/000711008X325300. PMid:18652737.
13. O'Hagan A, Stevens JW. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? Health Econ. 2003;12(1):33-49. http://dx.doi.org/10.1002/hec.699. PMid:12483759.
14. Jiang X, Guo X, Zhang N, Wang B, Zhang B. Robust multivariate nonparametric tests for detection of two-sample location shift in clinical trials. PLoS One. 2018;13(4):e0195894. http://dx.doi.org/10.1371/journal.pone.0195894. PMid:29672555.
15. Cleophas TJ. Clinical trials: robust tests are wonderful for imperfect data. Am J Ther. 2015;22(1):e1-5. http://dx.doi.org/10.1097/MJT.0b013e31824c3ee1. PMid:23896742.
16. Wagstaff DA, Elek E, Kulis S, Marsiglia F. Using a nonparametric bootstrap to obtain a confidence interval for Pearson's r with cluster randomized data: a case study. J Prim Prev. 2009;30(5):497-512. http://dx.doi.org/10.1007/s10935-009-0191-y. PMid:19685290.
17. Rascati KL, Smith MJ, Neilands T. Dealing with skewed data: an example using asthma-related costs of medicaid clients. Clin Ther. 2001;23(3):481-98. http://dx.doi.org/10.1016/S0149-2918(01)80052-7. PMid:11318082.
18. Vickers AJ, Altman DG. Statistics notes: missing outcomes in randomised trials. BMJ. 2013;346:1-4. http://dx.doi.org/10.1136/bmj.f3438. PMid:23744649.
19. Altman DG, Bland JM. Missing data. BMJ. 2007;334(7590):424. http://dx.doi.org/10.1136/bmj.38977.682025.2C. PMid:17322261.
20. Miot HA, Medeiros LM, Siqueira CRS, et al. Association between coronary artery disease and the diagonal earlobe and preauricular creases in men. An Bras Dermatol. 2006;81:29-33. http://dx.doi.org/10.1590/S0365-05962006000100003.
21. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. http://dx.doi.org/10.1136/bmj.b2393. PMid:19564179.
22. Little RJ. Regression with missing X's: A review. J Am Stat Assoc. 1992;87:1227-37.
23. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157-66. http://dx.doi.org/10.2147/CLEP.S129785. PMid:28352203.
24. Enders CK. Multiple imputation as a flexible tool for missing data handling in clinical research. Behav Res Ther. 2017;98:4-18. http://dx.doi.org/10.1016/j.brat.2016.11.008. PMid:27890222.
25. Stanimirova I, Walczak B. Classification of data with missing elements and outliers. Talanta. 2008;76(3):602-9. http://dx.doi.org/10.1016/j.talanta.2008.03.049. PMid:18585327.
26. Mackinnon A. The use and reporting of multiple imputation in medical research - a review. J Intern Med. 2010;268(6):586-93. http://dx.doi.org/10.1111/j.1365-2796.2010.02274.x. PMid:20831627.

27. Harel O, Mitchell EM, Perkins NJ, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):576-84. http://dx.doi.org/10.1093/aje/kwx349. PMid:29165547.

28. Enders CK. Multiple imputation as a flexible tool for missing data handling in clinical research. Behav Res Ther. 2017;98:4-18. http://dx.doi.org/10.1016/j.brat.2016.11.008. PMid:27890222.

29. Nunes LN, Klück MM, Fachel JMG. Multiple imputations for missing data: a simulation with epidemiological data. Cad Saude Publica. 2009;25(2):268-78. http://dx.doi.org/10.1590/S0102-311X2009000200005. PMid:19219234.

30. Miot HA. Correlation analysis in clinical and experimental studies. J Vasc Bras. 2018;17(4):275-9. http://dx.doi.org/10.1590/1677-5449.174118. PMid:30787944.

31. Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat Methods Med Res. 2018;27(9):2610-26. http://dx.doi.org/10.1177/0962280216683570. PMid:28034175.

32. Gades NM, Jacobson DJ, McGree ME, et al. Dropout in a longitudinal, cohort study of urologic disease in community men. BMC Med Res Methodol. 2006;6(1):58. http://dx.doi.org/10.1186/1471-2288-6-58. PMid:17169156.

33. Curran D, Molenberghs G, Aaronson NK, Fossa SD, Sylvester RJ. Analysing longitudinal continuous quality of life data with dropout. Stat Methods Med Res. 2002;11(1):5-23. http://dx.doi.org/10.1191/0962280202sm270ra. PMid:11923994.

34. Cheng J, Edwards LJ, Maldonado-Molina MM, Komro KA, Muller KE. Real longitudinal data analysis for real people: building a good enough mixed model. Stat Med. 2010;29(4):504-20. PMid:20013937.

35. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med. 2013;86(3):343-58. PMid:24058309.

36. Moreno-Betancur M, Chavance M. Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. Stat Methods Med Res. 2016;25(4):1471-89. http://dx.doi.org/10.1177/0962280213490014. PMid:23698867.

37. Rombach I, Jenkinson C, Gray AM, Murray DW, Rivero-Arias O. Comparison of statistical approaches for analyzing incomplete longitudinal patient-reported outcome data in randomized controlled trials. Patient Relat Outcome Meas. 2018;9:197-209. http://dx.doi.org/10.2147/PROM.S147790. PMid:29950913.

38. Garcia TP, Marder K. Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model. Curr Neurol Neurosci Rep. 2017;17(2):14. http://dx.doi.org/10.1007/s11910-017-0723-4. PMid:28229396.

39. Edwards LJ. Modern statistical techniques for the analysis of longitudinal data in biomedical research. Pediatr Pulmonol. 2000;30(4):330-44. http://dx.doi.org/10.1002/1099-0496(200010)30:4<330::AID-PPUL10>3.0.CO;2-D. PMid:11015135.

40. Miot HA. Survival analysis in clinical and experimental studies. J Vasc Bras. 2017;16:267-9. http://dx.doi.org/10.1590/1677-5449.001604. PMid:29930659.

41. Little R, Kang S. Intention-to-treat analysis with treatment discontinuation and missing data in clinical trials. Stat Med. 2015;34(16):2381-90. http://dx.doi.org/10.1002/sim.6352. PMid:25363683.

42. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ. 2011;342:1-9. http://dx.doi.org/10.1136/bmj.d40. PMid:21300711.

43. Streiner D, Geddes J. Intention to treat analysis in clinical trials when there are missing data. Evid Based Ment Health. 2001;4(3):70-1. http://dx.doi.org/10.1136/ebmh.4.3.70. PMid:12004740.

44. Bagatin E, Miot HA. How to design and write a clinical research protocol in Cosmetic Dermatology. An Bras Dermatol. 2013;88(1):69-75. http://dx.doi.org/10.1590/S0365-05962013000100008. PMid:23539006.

45. Resseguier N, Giorgi R, Paoletti X. Sensitivity analysis when data are missing not-at-random. Epidemiology. 2011;22(2):282. http://dx.doi.org/10.1097/EDE.0b013e318209dec7. PMid:21293212.

46. Yamaguchi Y, Misumi T, Maruo K. A comparison of multiple imputation methods for incomplete longitudinal binary data. J Biopharm Stat. 2018;28(4):645-67. http://dx.doi.org/10.1080/10543406.2017.1372772. PMid:28886277.

47. Wen L, Terrera GM, Seaman SR. Methods for handling longitudinal outcome processes truncated by dropout and death. Biostatistics. 2018;19(4):407-25. http://dx.doi.org/10.1093/biostatistics/kxx045. PMid:29028922.

48. Spratt M, Carpenter J, Sterne JA, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. 2010;172(4):478-87. http://dx.doi.org/10.1093/aje/kwq137. PMid:20616200.

49. Ferretti F, Saltelli A, Tarantola S. Trends in sensitivity analysis practice in the last decade. Sci Total Environ. 2016;568:666-70. http://dx.doi.org/10.1016/j.scitotenv.2016.02.133. PMid:26934843.

50. Tseng CH, Elashoff R, Li N, Li G. Longitudinal data analysis with non-ignorable missing data. Stat Methods Med Res. 2016;25(1):205-20. http://dx.doi.org/10.1177/0962280212448721. PMid:22637472.

Correspondência
Hélio Amante Miot
Universidade Estadual Paulista – UNESP
Av. Prof. Mário Rubens Guimarães Montenegro, s/n - Distrito de Rubião Junior
CEP 18618-687 - Botucatu (SP), Brasil
Tel.: (14) 3882-4922
E-mail: heliomiot@gmail.com

Informações sobre os autores
HAM - Livre-docente, Departamento de Dermatologia e Radioterapia da Faculdade de Medicina de Botucatu, Universidade Estadual Paulista (UNESP), Botucatu, SP, Brasil.