Journal of Statistics on Health Decision

EXTENDED ABSTRACT

P30

Single versus Multiple Imputation Methods Applied to Classify Dyslipidemic Patients Concerning Statin Usage: a Comparative Performance Study

João Albuquerque[1,2], Ana C. Alves[3,4], Ana M. Medeiros[3,4], Mafalda Bourbon[3,4], Marília Antunes[1,5]

[1] Centro de Estatística e Aplicações da Universidade de Lisboa, FCUL, Portugal
[2] Departamento de Bioquímica, FMUP, Portugal
[3] Cardiovascular Research Group, INSA, Lisboa, Portugal
[4] Instituto de Biossistemas e Ciências Integrativas – BioISI, FCUL, Portugal.
[5] Departamento de Estatística e Investigação Operacional, FCUL, Portugal

Introduction

One of the greatest challenges when working with clinical datasets is to decide how to deal with missing values. Removing observations with any missing values prior to data analysis, a process defined as listwise deletion, is the standard default procedure in most statistical software packages, but may lead to great loss of valuable information [1]. The use of robust imputation methods may provide accurate estimates for missing values, allowing to include these observations into the analysis. The imputation strategy to adopt depends on the amount and type of missing information, and also on the relation between variables, allying statistical expertise with clinical understanding of the data. The main purpose of this work was to compare the performance of two different methods of imputation to overcome missingness on dyslipidemic patients regarding statin usage.

Methods

The sample used in this study was constituted by 512 adult participants of the Portuguese Familial Hypercholesterolemia (FH) Study, of both sexes (mean age=45, range 18-78 years). The dataset was constituted by 28 predictor variables, which included personal, clinical, and biochemical information, plus the dependent variable, a categorical binary outcome, indicating statin usage (Y/N). Nine of the predictor variables presented missing values. The proportion of missing values for five of these variables was relatively low (2-12%). The remaining four variables, concerning total cholesterol (TC), low (LDLc) and high (HDLc) density lipoprotein cholesterol, and triglycerides plasmatic concentrations, assessed in a moment previous to first medical consultation, presented high proportion of missing values (29-38%). These variables are considered very important predictors of statin usage when present, since they allow calculating the respective percentage of variation (perc.red) for different cholesterol subfractions, between that moment and date of last consultation, when the patient may be medicated or not. The dependent variable presented a fraction of missing values around 15%. A single and a multiple imputation methods were used for performance comparison. In the first case, the k nearest neighbor (kNN) method based on Gower similarity measure [2] was used to impute missing values in predictor variables with low proportion of missing values, and a logistic regression (LR) model was developed to predict statin usage. Due to high percentage of missing values, perc.red values were not imputed in this approach. Instead, two separate LR models were built, LR1, for cases in which these variables were present (n=312), and LR2, using the complete dataset, and excluding perc.red variables. In the second method, the multiple imputation by chained equations (MICE) algorithm was applied to the entire dataset, adopting a number of imputed datasets m=5 [1]. Internal validation was performed for every model, through 10-fold cross validation (CV), and agreement between both methods in predicting statin usage was calculated.

Results

The final LR1 model included the variables perc.red LDLc, perc.red TC, Physical Signs and Hypertension, by order of importance. 10-fold CV revealed a mean accuracy (Acc) level of 0.967, suggesting very

Keywords:
Data imputation; Statins; Dyslipidemia

Corresponding author:
João Albuquerque
joaodavid.alb@gmail.com

Conflict of interest:
The authors declare no conflict of interests

First published: 23 OCT 2020

Open Access Publication

© 2020 Albuquerque J, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

good predictive ability for this model. The final LR2 model, which did not include perc.red variables, retained the variables LDLc, Age, FH diagnosis, Body Mass Index, Physical Signs, Hypertension and Lipoprotein(a), by order of importance. 10-fold CV showed a lower mean Acc level of 0.854. Finally, the MICE model included the variables perc.red LDLc, perc.red TC and Physical Signs, and presented mean Acc levels ranging from 0.933-0.947. Agreement between LR1 and MICE was of 0.8, between LR2 and MICE of 0.725 and overall between both LR models and MICE of 0.775.

Discussion and conclusions

The predictive ability of LR models to impute statin usage seems to be greatly affected by the availability of information reflecting variation in cholesterol levels. Even with great percentage of missing values concerning this information, MICE method seems to maintain quality of imputation, and may be considered a valid alternative in such cases. When applied to a new set of unclassified data, agreement between both methods seems to be limited. In order to decide which method behaves better with previously unseen data, a testing set of pre-classified subjects should be created, and incorporated in the analysis.

Acknowledgements

Research supported by the programme Norte2020 (operação NORTE-08-5369-FSE-000018) and by national FCT funds under the projects UID/MAT/00006/2019 and PTDC/SAU-SER/29180/2017.

References

1. Van Buuren, S. Flexible imputation of missing data. CRC press. 2018
2. Kowarik, A; Templ, M. Imputation with the R Package VIM. Journal of Statistical Software. 2016;74(7):1-16. https://doi.org/10.18637/jss.v074.i07