Machine Learning Methods to Handle Missing PHQ-8 Score Values in the UAEHFS Pilot Data

Mitha Al Balushi1, Amar Ahmad1, Syed Javaid2, Luai Ahmed2, Fatma Al-Maskari2, Abdishakur Abdulle1, Raghib Ali1

1 Public Health Research Center, New York University Abu Dhabi, United Arab Emirates; 2 College of Medicine and Health Sciences, United Arab Emirates University, United Arab Emirates

INTRODUCTION

The UAE Healthy Future Study (UAEHFS) is one of the first large prospective cohort studies in the region which examines causes and risk factors for chronic diseases among adult UAE nationals. Missing values are often unavoidable in empirical research and can lead, in many cases, to bias when missing data are omitted in the statistical analysis. The eight-item Patient Health Questionnaire (PHQ-8) is one of the important variables included in the UAEHFS which are collected with missing values.

OBJECTIVES

The aim of this study was to estimate the percentage of reported depression using different statistical machine learning methods of handling missing values using the UAEHFS pilot data.

METHODS

Complete case was included in the primary analysis. In a sensitivity analysis, five common statistical machine learning methods of handling missing values are included in the analysis. These five methods are mode imputation, k-nearest neighbour (KNN) imputation, classification, and regression trees (CART), random forest (RF) imputations, and random sample from observed values (Sample).

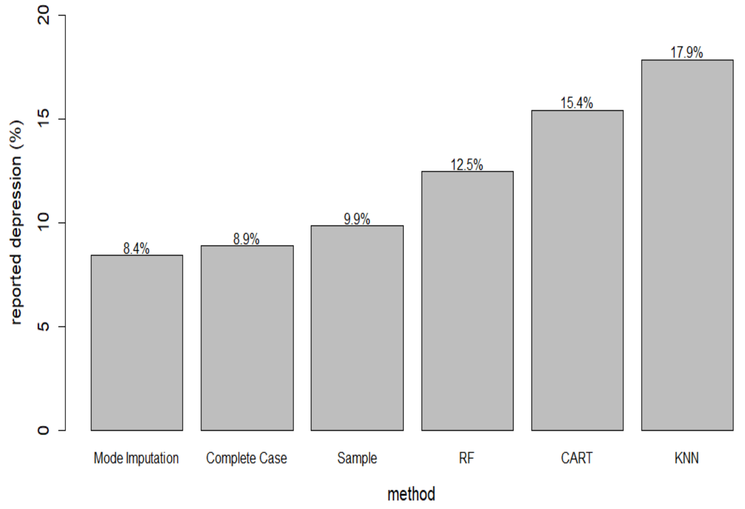
RESULTS

Out of 517 participants, data from 487 (94.2%) were analyzed after excluding participants who didn't fill out the questionnaires. The median age was 30 years (Interquartile Range: 23 - 38). There were more males (67.8%) than females in the UAEHFS pilot data. The pattern of missing values was investigated, and it was found that subjects who "did not want to answer" were not systematically different (in term of age and gender) from those who answered the questionnaire. Therefore, missing at random (MAR) was assumed. The estimated percentage of reported depression was 8.4%, 8.9%, 9.9%, 12.5%, 15.4% and 17.9% by the mode imputation, complete case, sample, RF, CART, and KNN respectively.

CONCLUSIONS

The estimated percentage of reported depression varies between the six applied statistical machine learning approaches. This shows that the problem of missing values in the variables is not negligible and is so common that it needs to be continuously studied and investigated. Further research is needed to address the issue of missing values using the main UAEHFS dataset after completing recruitment.

Table 1: estimated percentage of reported depression by different methods of handling missing values



REFERENCES

Abdulle A, Alnaeemi A, Aljunaibi A, Al Ali A, Al Saedi K, Al Zaabi E, Oumeziane N, Al Bastaki M, Al-Houqani M, Al Maskari F, Al Dhaheri A. The UAE healthy future study: a pilot for a prospective cohort study of 20,000 United Arab Emirates nationals. BMC Public Health. 2018 Dec;18(1):1-9

The Authors declare that there is no conflict of interest.

Contact: Ms Mitha Al Balushi, ma4643@nyu.edu