The First Scientific Conference on Health and Medical Research in the UAE-5-6 December 2022

A Statistical Machine Learning Method to Handle Missing PHQ-8 Score Data in the UAEHFS Pilot Data – A Bayesian Network Approach

Al Balushi M [1,3], Ahmad A [1] , Javaid S [2] , Ahmed L [3] , Grivna M [3] , Al Maskari F [3] , Abdulle A [1] , Ali R [1,4]

[1] Public health research center-New York University in Abu Dhabi, United Arab Emirates
[2] Department of Psychiatry and Behavioral Sciences, United Arab Emirates University, United Arab Emirates.
[3] Institute of public Health, College of Medicine and Health Sciences, United Arab Emirates University, United Arab Emirates
[4] MRC Epidemiology Unit, University of Cambridge, Oxford London, United Kingdom

Email: Mitha Al Balushi (ma4643@nyu.edu).

Abstract

Background and aims: 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. The aim of this study was to estimate the effect of body fat percentage (BF%) on self-reported depression using a Bayesian statistical machine learning approach of handling missing values using the UAEHFS pilot data.

Methods: Complete case was included in the primary analysis. In a sensitivity analysis, missing values in the UAEHFS pilot data were imputed from a Bayesian network. Ties in prediction are broken using sampling at random from the tied values. Therefore, setting the random seed is required to get reproducible results. A logistic regression model was performed with the dichotomized PHQ-8 variable as predictor and age, gender, BF% as well as the interaction terms of age*gender and gender*BF% as predictors. Odds ratios (ORs) with 95% confidence intervals (95% CI) and corresponding p-value was computed.

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. The estimated OR (95% CI) of BF% was 1.362 (1.083, 1.713), p-value = 0.008 and 1.190 (1.033, 1.371), p-value= 0.016 from the complete case and sensitivity analysis respectively.

Conclusions: The result of this study suggests that imputation of missing values reduces the amount of overfitting in the result. 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. The result of this study suggests that imputation of missing values reduces the amount of overfitting in the result. 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.