Cluster analysis on emergency COVID-19 data: A result-based multiple imputation for missing data

Halehsadat Nekoee Zahraei[1,4*], Allison Gilbert[2], Anh Nguyet Diep[1], Renaud Louis[4], Alexandre Ghuysen[2,3], Anne-Françoise Donneau[1]

[1] Biostatistics Unit, Department of Public Health, University of Liège, Belgium
[2]Emergency Department, University Hospital Center of Liège, Liège, Belgium
[3] Medical Simulation Center' Public Health Department, University of Liège, Liège, Belgium
[4]Department of Pneumology, GIGA, University of Liège, Belgium
* presenting author

Background and Objective

In 2020, hospitals have been confronted with an influx of COVID-19 confirmed patients. Grouping patients based on clinical features could help clinicians to identify a structure of patients who needs more attention. The present study considers cluster analysis to identify different clinical phenotypes with similar properties while accounting for the presence of missing data. Although several frameworks exist for handling missing data in cluster analysis, in this study, a new perspective was introduced for multiple imputation in cluster analysis that focused on the result of clustering.

Method

To handle the uncertainty of missing values, *m* imputed datasets were generated. The model-based clustering strategy was applied on the imputed datasets. Based on BIC criterion, the best method and the best number of groups were defined for all imputed datasets. Subsequently, the most repetitive number of groups and types was fixed. In the next step, cluster analysis was re-applied on *m* imputed datasets by the fixed number of clusters and type. The results of the statistical analysis were reported for each of the groups in imputed datasets. According to Rubin’s rules, in the pooled step, the final results were combined by mean and the statistical inferences were applied by considering between and within variance.

Results

The performance of the proposed framework was compared and assessed in several scenarios. The proposed method with 20 clinical features was performed on 628 confirmed COVID-19 patients who presented at University Hospital of Liege from March to May 2020. Based on model-based clustering and BIC criterion for multiple imputation, the patients were classified into four clusters. The rate of hospitalization in Cluster2 with older patients was higher than those in Cluster1. The oldest patients were assigned to Cluster3 and Cluster 4. The rate of comorbidity was almost close to 100% in Cluster 4 and percentage of infectious disease in cluster3 was less than Cluster4; however, Cluster3 had a higher rate of hospitalization than Cluster4.

Conclusions

The proposed method handled cluster analysis on missing data by multiple imputations. Also, the present study identified four clusters of patients confirmed with COVID-19 and the corresponding rate of hospitalization based on clinical features.