# COVID-19 patient profiles in the Basque Country: A clustering approach

*Lander Rodriguez*[1], *Irantzu Barrio*[2], *Daniel Fernández*[3], *José M. Quintana-Lopez*[4]

[1]lrodriguez@bcamath.org, Applied Statistics, Basque Center for Applied Mathematics

[2]irantzu.barrio@ehu.eus, Department of Mathematics, University of the Basque Country

[3]daniel.fernandez.martinez@upc.edu, Department of Statistics and Operations, Universitat Politecnica de Catalunya · BarcelonaTech

[4]josemaria.quintanalopez@osakidetza.eus, Research Unit of the Galdakao-Usansolo University Hospital, Osakidetza Basque Health Service

The ever-increasing availability of medical data has opened the way for Machine Learning and advanced statistical methods to be applied in health. In particular, clustering techniques discover hidden and inherent patterns to organize data into groups without any a priori hypothesis. This feature of clustering techniques can be used for patient profile identification, which would be advantageous for multiple care intervention strategies and to offer an improved medical attention. In this work, we create COVID-19 positives patient profiles from a population-based database based on the novel KAMILA clustering technique (Foss et al., 2016). This technique overcomes the problems faced by clustering methods when dealing with mixed-type data, which is often the case in clinical research. In addition, it is appropriate for large datasets, which is the case in this work.

All the patients included in this study were residents in the Basque Country and were diagnosed COVID-19 from March 1, 2020 until January 9, 2022. The data included sociodemographic data, baseline comorbidities and baseline treatments. In addition, COVID-19 adverse outcomes were also included: hospitalization, adverse evolution (ICU or death) and death. A two-stage process was implemented: first we identified the profiles of COVID-19 positives in the Basque Country and then we assessed their association with the adverse outcomes of the disease. The profiles were created for different periods with the KAMILA clustering technique and their evolution in time was assessed.

Age and the Charlson index were the variables that mainly differentiated the profiles, together with, but to a lesser extent, diabetes, kidney disease, metastatic solid tumor and heart failure. The patient profiles were well differentiated by their risk to the adverse outcomes of COVID-19. Actually, the severity of the outcomes increased with the risk level of the clusters for all the periods and outcomes. Apart from that, the profiles evolved to lower risk profiles along the pandemic, which was reflected in the COVID-19 hospitalization, adverse evolution and death reduction.

To our best knowledge, this is the first study used to create COVID-19 patient profiles from COVID-19 positives of the population and to assess their evolution in time. The previous results suggest the appropriateness of clustering techniques and particularly KAMILA to identify risk profiles in large electronic health records with mixed-type data. This could lead to a better allocation of the health resources and an improved medical attention.

**Keywords**: COVID-19, clustering.