Socioeconomic factors in predictive models:

Understanding COVID-19 in Brazil, and in other highly unequal societies

Valentina Martufi[1,2], Renzo Flores-Ortiz[1], Priscilla Normando[1], Vinicius A. Oliveira[1], Maria Yury Ichihara[1], Mauricio L. Barreto[1,2], Elzo P. P. Júnior[1]

[1] Center for Data and Knowledge Integration for Health (CIDACS), Gonçalo Moniz Institute, Oswaldo Cruz Foundation (FIOCRUZ), Salvador, Brazil

[2] Institute of Collective Health, Federal University of Bahia, Salvador, Brazil

BACKGROUND

The Covid-19 pandemic provided the motive and momentum for the development of internationally coordinated efforts to analyse health data to be able to understand the disease's manifestation (symptoms), risk factors (including co-morbidities), and potential outcomes (morbimortality) [1]. Nonetheless, this global health emergency soon manifested its socio-economic facets, as we saw the most disadvantaged communities being affected exponentially more than the wealthier sections of the population [2] [3], as well as becoming apparent that the Covid-19 pandemic was capable of exacerbating even further already existing socio-economic inequalities [4] [5]. Subsequently, it is important to have a comprehensive understanding of all of the factors that affect and are affected by any given globally significant health-related event, such as Covid-19, to promote proper care provision and prescription of treatment at the individual health level, as well as improving the overall use of medical resources [6] and incentivizing better informed policy-making, planning and organization of work processes and inter-sectorial collaboration at the public health level [5]. In this regard, public sector, administrative datasets, containing what is increasingly being referred to as "real-world evidence", represent an unprecedented resource to develop socially significant and appliable research. In this study, we aimed to use real-world data to build bio-socio-economic prediction models for mortality and intensive care unit (ICU) admission among patients hospitalized with suspected COVID-19 in Brazil, to support managers and policymakers in dealing with a phenomenon such as the COVID-19 pandemic.

METHODS

This is a retrospective cohort study of patients hospitalised with COVID-19 between February 2020 and July 2022 in Brazil. The study conduct and reporting of methodology and findings followed the Transparent Reporting of a multivariate prediction model for Individual Prediction or Diagnosis (TRIPOD) guidelines [7]

Patients were followed from the date of hospital admission to the date of discharge or death. All enrolled patients were over 18 years old and were tested for COVID-19 by RT-PCR or rapid antigen testing. The patients' data were retrieved from the *Síndrome Respiratória Aguda Grave* (SRAG) database [8], which includes national records of hospitalizations due to Severe Acute Respiratory Syndrome (SARS), and is made openly available by the Ministry of Health.

The study outcome is a composite variable of the occurrence of at least one of three critical in-hospital events: invasive mechanical ventilation (IMV) support, intensive-care unit (ICU) admission or death. After reviewing the existing literature on risk factors of critical in-hospital events among COVID-19 patients [6] [9] we selected 31 variables from the SRAG database to perform predictive modelling, namely: age, sex, race, education level (later excluded because of excessive missing values – 61.9%), material deprivation (measured with the Brazilian Deprivation Index [10]) as a proxy of socioeconomic status, macro-region of residence, pre-existing comorbidities (cardiovascular disease, diabetes, obesity, cancer, asthma, immunodeficiency, chronic kidney disease, other chronic lung disease, chronic hematologic disease, down syndrome, chronic liver disease, chronic neurological disease), symptoms of severe acute respiratory syndrome (fever, cough, sore throat, dyspnoea, respiratory distress, low oxygen saturation, diarrhoea, vomit, abdominal pain, fatigue, loss of smell, loss of taste), and an indicator of vaccination against COVID-19.

The data were randomly split for the model building (training set: 75%) and model validation (test set: 25%). We then balanced the number of outcome events in the training set to ensure correct predictions. Model parameters were estimated on the training set conditional on the value(s) of the tuning parameters. Model validation was applied to the test set. In total numbers, 35,616 patients were used for model training and 19,484 for validation.

We used logistic regression and machine learning approaches to build prediction models of critical in-hospital events. We first estimated a logistic regression model with all predictors and dropped those that were not statistically significant. The remaining predictors were then used to build prediction models of critical in-hospital events using logistic regression and machine learning approaches. We evaluated the predictive performance of the models using the AUC and the root mean square of residuals. We assessed the relevance of the predictors to the model predictions using the loss of variable importance (1-AUC) after permutation [11]. Analyses were performed using the statistical software R version 3.6.

RESULTS

In total, 97,768 patients were enrolled in the study cohort. During 1,163,089 person-days, 36,358 (37.2%) ICU admissions, 67,078 (68.6%) received mechanical ventilation support, and 25,775 (26.4%) deaths occurred. Overall, 75,457 patients (77.2%) had experienced at least one critical in-hospital event. At the date of hospital admission, the median age was 61 years (interquartile range 46, 74 years), with most men (54.5%) and individuals of white race (49.6%), predominantly resident in the South (23.8%) or Southeast (49.6%) regions. The median hospitalization duration was 8 days (interquartile range 4, 14 days). The most prevalent chronic conditions were cardiovascular disease (44.0%), diabetes (23.9%), and obesity (8.0%), while the least prevalent was hematologic disease (0.8%). The most prevalent COVID-19 symptoms were cough (69.5%), dyspnoea (68.4%), and fever (55.9%), while the least prevalent was abdominal pain (0.8%).

Overall, the models exhibit little variation in performance for critical in-hospital events prediction (figure 1). AUC values ranged from 0.6515 to 0.7002 (highest for the neural network model) and root mean square of residuals from 0.468 to 0.563 (lowest for the neural network model) (figure 2).

In the neural network model (figure 3), the fifth and eleventh most relevant predicting factors for our outcome were, respectively, region of residence and socioeconomic deprivation level. The former featured right after respiratory distress and before having been vaccinated against Covid-19, whilst the latter appeared to be more important than suffering from immunodeficiency, cardiovascular disease and other chronic lung disease.

CONCLUSIONS

The results of this study highlight the relevance of including socioeconomic factors in modelling efforts to understand Covid-19 or any other health phenomena, if the purpose is to inform properly coordinated intersectoral actions to preserve the health of the population.

REFERENCES

[1] Elzo Pereira Pinto Junior, Priscilla Normando, Renzo Flores-Ortiz, Muhammad Usman Afzal, Muhammad Asaad Jamil, Sergio Fernandez Bertolin, Vinícius de Araújo Oliveira, Valentina Martufi, Fernanda de Sousa, Amir Bashir, Edward Burn, Maria Yury Ichihara, Maurício L Barreto, Talita Duarte Salles, Daniel Prieto-Alhambra, Haroon Hafeez, Sara Khalid. **Integrating real-world data from Brazil and Pakistan into the OMOP common data model and standardized health analytics framework to characterize COVID-19 in the Global South**. *Journal of the American Medical Informatics Association*, Volume 30, Issue 4, April 2023, Pages 643–655, DOI: https://doi.org/10.1093/jamia/ocac180

[2] Elissa M Abrams & Stanley J Szefler. **COVID-19 and the impact of social determinants of health.** *The Lancet Respiratory Medicine*, COMMENT| VOLUME 8, ISSUE 7, P659-661, JULY 2020. DOI: https://doi.org/10.1016/S2213-2600(20)30234-4

[3] Sage J. Kim & Wendy Bostwick. **Social Vulnerability and Racial Inequality in COVID-19 Deaths in Chicago**. *Health Education & Behavior* 2020, Vol. 47(4) 509–513. DOI: https://doi.org/10.1177/1090198120929677

[4] Sean A.P. Clouston, Ginny Natale, Bruce G. Link. **Socioeconomic inequalities in the spread of coronavirus-19 in the United States: A examination of the emergence of social inequalities.** *Social Science & Medicine*, Volume 268, 2021, 113554, ISSN 0277-9536, DOI: https://doi.org/10.1016/j.socscimed.2020.113554.

[5] Karina Braga Ribeiro, Ana Freitas Ribeiro, Maria Amélia de Sousa Mascena Veras, Marcia Caldas de Castro. **Social inequalities and COVID-19 mortality in the city of São Paulo, Brazil.** *International Journal of Epidemiology*, Volume 50, Issue 3, June 2021, Pages 732–742, DOI: https://doi.org/10.1093/ije/dyab022

[6] Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, et al. **Prediction for Progression Risk in Patients With COVID-19 Pneumonia: The CALL Score.** *Clin Infect Dis.* 2020;71(6):1393–9.

[7] Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;**13**:1. doi:10.1186/s12916-014-0241-z

[8] Ministério da Saúde. **Banco de Dados de Síndrome Respiratória Aguda Grave - SRAG** [Internet]. Available from: https://opendatasus.saude.gov.br/dataset/srag-2020

[9] Gopalan N, Senthil S, Prabakar NL, Senguttuvan T, Bhaskar A, Jagannathan M, et al. **Predictors of mortality among hospitalized COVID-19 patients and risk score formulation for prioritizing tertiary care—An experience from South India.** Dhali GK, editor. *PLoS One*. 2022;17(2):e0263471.

[10] CIDACS-FIOCRUZ. **IBP - Índice Brasileiro de Privação** [Internet]. [cited 2021 Oct 2]. Available from: https://cidacs.bahia.fiocruz.br/ibp/

[11] Biecek P, Burzykowski T. Explanatory model analysis: explore, explain, and examine predictive models. CRC Press 2021.

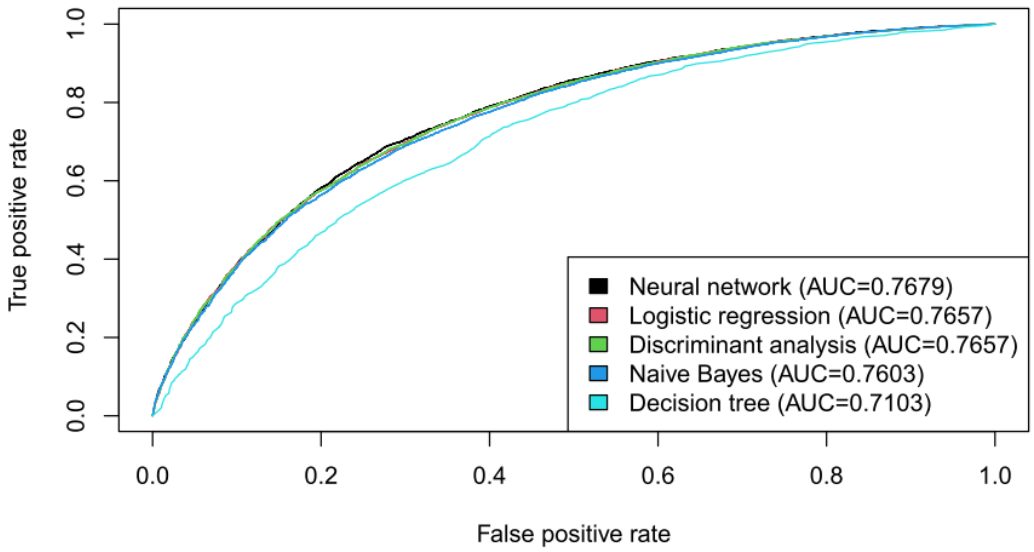

Figure 1. Receiver operating characteristic curves and AUC values of critical in-hospital events among patients hospitalised with laboratory-confirmed COVID-19 in Brazil. AUC, area under the receiver operating characteristic curve.

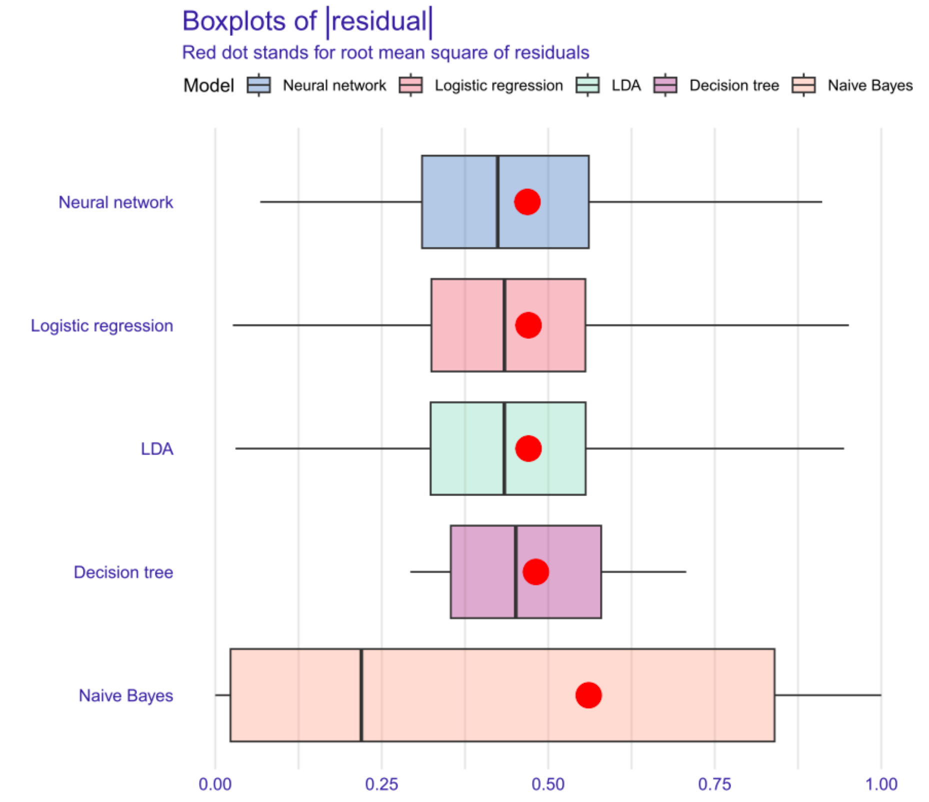

Figure 2. Boxplots of residuals for predicting critical in-hospital events among patients hospitalised with confirmed COVID-19 in Brazil.

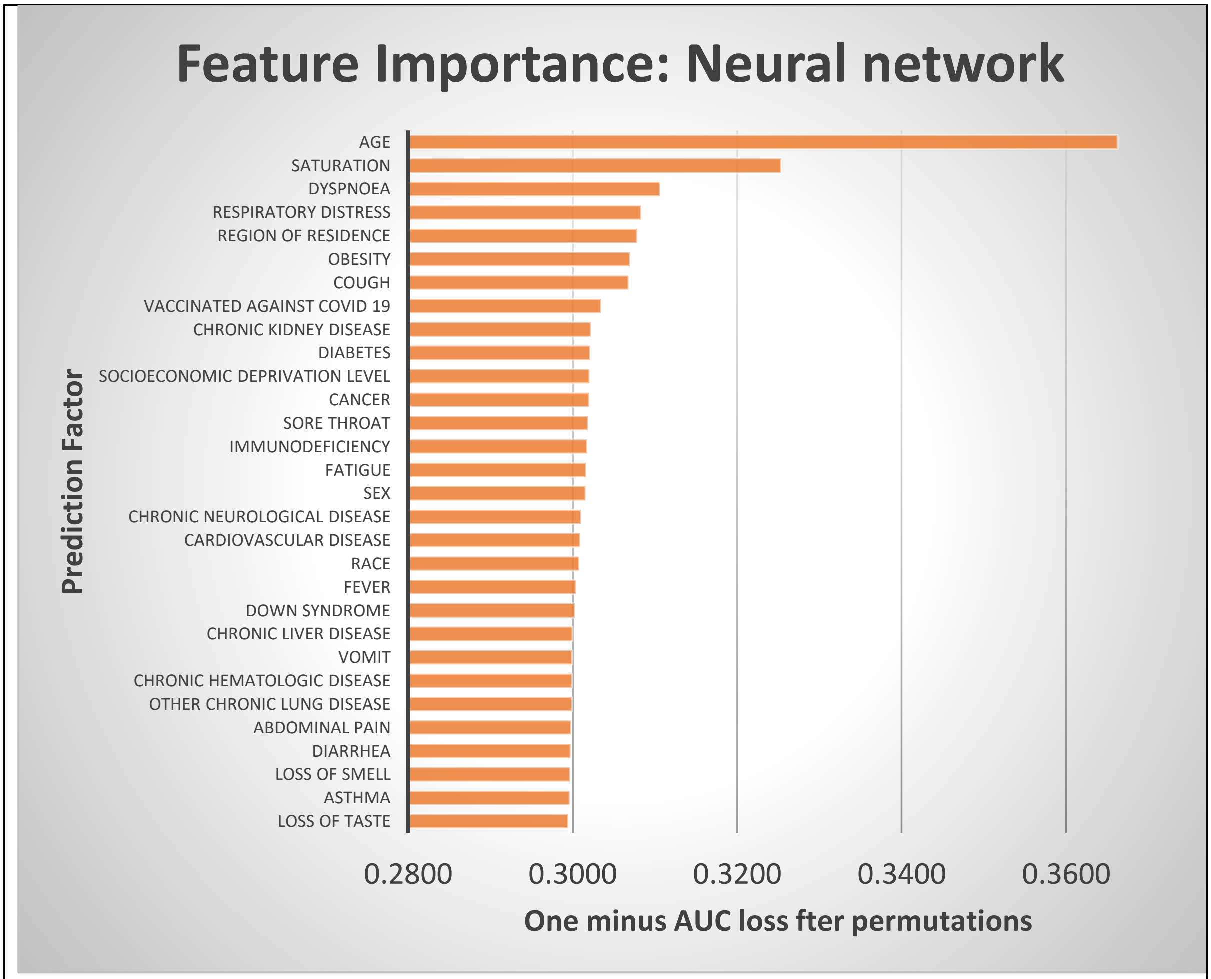

Figure 3. Importance of predictor variables in the neural network model predicting critical in-hospital events among patients hospitalised with COVID-19 in Brazil.