



OPEN

A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil

Fernando Timoteo Fernandes^{1,2✉}, Tiago Almeida de Oliveira^{1,3},
Cristiane Esteves Teixeira^{1,4}, Andre Filipe de Moraes Batista¹, Gabriel Dalla Costa⁵ &
Alexandre Dias Porto Chiavegatto Filho¹

The new coronavirus disease (COVID-19) is a challenge for clinical decision-making and the effective allocation of healthcare resources. An accurate prognostic assessment is necessary to improve survival of patients, especially in developing countries. This study proposes to predict the risk of developing critical conditions in COVID-19 patients by training multipurpose algorithms. We followed a total of 1040 patients with a positive RT-PCR diagnosis for COVID-19 from a large hospital from São Paulo, Brazil, from March to June 2020, of which 288 (28%) presented a severe prognosis, i.e. Intensive Care Unit (ICU) admission, use of mechanical ventilation or death. We used routinely-collected laboratory, clinical and demographic data to train five machine learning algorithms (artificial neural networks, extra trees, random forests, catboost, and extreme gradient boosting). We used a random sample of 70% of patients to train the algorithms and 30% were left for performance assessment, simulating new unseen data. In order to assess if the algorithms could capture general severe prognostic patterns, each model was trained by combining two out of three outcomes to predict the other. All algorithms presented very high predictive performance (average AUROC of 0.92, sensitivity of 0.92, and specificity of 0.82). The three most important variables for the multipurpose algorithms were ratio of lymphocyte per C-reactive protein, C-reactive protein and Braden Scale. The results highlight the possibility that machine learning algorithms are able to predict unspecific negative COVID-19 outcomes from routinely-collected data.

The consequences of a long stay and demand for hospital resources due to COVID-19 have been disastrous for health systems in middle and low-income countries (LMICs)^{1,2}, requiring immediate clinical decisions, especially when dealing with limited resources^{3,4}. An accurate COVID-19 prognosis assessment is crucial for screening and treatment procedures and may increase patient survival^{5,6}. In Brazil⁷, many cities are at their saturation capacity for the provision of clinical care, especially regarding ICU beds and mechanical ventilators^{8–20}. Data-driven solutions are needed to support decision-making¹¹.

COVID-19 has shown to rapidly worsen a few days after infection^{12,13}. The median time from disease onset to ICU admission is 9–12 days^{14,15}. About 26–32% of the hospitalized patients are eventually admitted to ICU, and mortality in this group ranges from 39 to 72%, depending on the local characteristics of patients^{14,15}. The median length of ICU stay and use of mechanical ventilation is approximately 9 days (95% CI 6.5–11.2) and 8.4 days (95% CI 1.6–13.7), respectively¹⁶.

Previous studies have used blood tests¹⁷, CT images^{18,19}, sociodemographic and comorbidities history²⁰ to develop COVID-19 diagnostic and prognostic models, including machine learning techniques^{21–23}. Biomarkers from blood tests have emerged as important variables for poor prognostic factors²⁴, which are a promising tool in poorer regions, due to its low cost and inclusion in standard protocols for clinical care. However, the majority of studies²⁵ rely on algorithms trained on a single prognostic outcome, which in theory require the training of specific algorithms for each distinct negative outcome.

¹School of Public Health, University of São Paulo, São Paulo, SP, Brazil. ²Fundacentro, São Paulo, SP, Brazil. ³Statistics Department, Paraíba State University, Paraíba, PB, Brazil. ⁴Bioinformatics and Computational Biology Lab, Brazilian National Cancer Institute, Rio de Janeiro, RJ, Brazil. ⁵BP-A Beneficência Portuguesa de São Paulo, São Paulo, SP, Brazil. ✉email: fernando.fernandes@fundacentro.gov.br