ISARIC (International Severe Acute Respiratory and Emerging Infections Consortium)

A global federation of clinical research networks, providing a proficient, coordinated, and agile research response to outbreak-prone infectious disease

Analysis Plan for ISARIC International COVID-19 Patients

Title of proposed research
Statistical and machine learning methods for predicting risk of pulmonary embolism and death in patients with COVID-19
Version: (Date: Day/Month/Year)
22/3/2022
Working Group Chair (name, ORCID ID, email, institution, country)
Christiana Kartsonaki (christiana.kartsonaki@dph.ox.ac.uk, University of Oxford, ORCID: 0000-0002-3981-3418)
[1]Working group co-chair (name, ORCID ID, email, institution, country)
Lei Clifton (lei.clifton@ndph.ox.ac.uk, University of Oxford) Munib Mesinovic (munib.mesinovic@jesus.ox.ac.uk, University of Oxford) Giri Shan Rajahram, gsrajahram@gmail.com, Queen Elizabeth Hospital, ORCID: 0000-0002-8123-476X Wong Xin Ci, wongxc.crc@gmail.com, Digital Health Research and Innovation Unit, Malaysia, ORCID: 0000-0002-1036-8023
Statistician (name, ORCID ID, email, institution, country)
Munib Mesinovic (munib.mesinovic@jesus.ox.ac.uk, University of Oxford)

[1] Either chair and/or co-chair are based in an institution in an LMIC. If you would like to be connected with an eligible co-chair please let us know at ncov@isaric.org.

Final draft SAPs will be circulated to all ISARIC partners for their input with an invitation to participate. ISARIC can help to set up collaborator meetings; form a working group; support communications; and accessing data. Please note that the details of all approved applications will be made publicly available on the ISARIC website. Please complete all sections of this form fully and return to ncov@isaric.org

Introduction

Hospitalised patients with COVID-19 are at risk of a number of complications, such as pulmonary embolism, which early in the pandemic had an incidence of 2.6–8.9% among all hospitalised patients and up to one third of patients requiring intensive care (Sakr et al., 2020).

The aim of the project is to explore methods for predicting clinical outcomes (e.g. in-hospital pulmonary embolism and mortality) in cohorts of hospitalised patients and compare their properties, with application to data from a cohort of hospitalised patients with COVID-19.

This project will investigate both classical statistical models (e.g. logistic and time-to-event regression with and without regularization) and machine learning approaches (e.g. tree-based algorithms such as extreme gradient boost machine, XGBoost). These methods will be used to predict outcomes using a large number of clinical variables (including demographics, comorbidities, and risk factors) on patients hospitalised with COVID-19. The main objective will be to review, apply, and compare the properties of different methods, with application to data from ISARIC.

This is a student project for the Centre for Doctoral Training in Health Data Science at the University of Oxford.

This document details the initial analysis plan for publication on a subset of COVID-19 patients in the global cohort in the ISARIC database, as of January 2022. There are currently 53 countries contributing data and these have so far contributed data on ~800,000 patients. These data will represent the global experience of the first 2 years of this pandemic.

Participatory Approach

All contributors to the ISARIC database are invited to participate in this analysis through review and input on the statistical analysis plan and resulting publication. The outputs of this work will be disseminated as widely as possible to inform patient care and public health policy, this will include submission for publication in an international, peer-reviewed journal. ISARIC aims to include the names of all those who contribute data as cited collaborators of this publication, subject to the submission of contact details and confirmation of acceptance of the final manuscript within the required timelines, per ICMJE policies and the ISARIC publication policy.

Research Plan

Summary of Research Objectives

To investigate and compare methods for predicting risk of pulmonary embolism and death among patients with COVID-19

Proposed Target Population

Hospitalised patients with COVID-19

Clinical Questions/Descriptive Analyses

Can the risk of pulmonary embolism (and of [any] death) during hospitalisation with COVID-19 be predicted by patient characteristics at admission?

How well do various approaches for predicting risk of a binary (and time-to-event) outcome perform in this setting?

The project will involve:

- Conducting a brief literature review.
- Implement statistical and tree-based machine learning methods on COVID-19 data.
- Comparison of the methods and results.

This project can act as a foundation for a DPhil, which may include:

- Comparison of methods for predicting outcomes, applied to data and potentially using simulation studies.
- Extensions to deep learning for time-to-event outcomes.
- Comparisons of measures for evaluating model performance.

Data from ISARIC will be used. If expanded into a DPhil project, data from the UK Biobank may also be used.

Planned Statistical Analyses, Methodology and Representation

1. Summary statistics of key demographic variables.
2. Numbers of patients who developed pulmonary embolism during hospitalization.
3. Summary statistics of symptoms at presentation, comorbidities, risk factors and lab results.
4. Fitting different models for predicting risk of pulmonary embolism given patient characteristics at admission (age, sex, country, symptoms, comorbidities, risk factors, lab results):
   - Logistic regression
   - Logistic regression with regularization (e.g. lasso, ridge, elastic net)
   - Logistic regression with regularization including interaction terms
   - Random forests

- Extreme gradient boost machine, XGBoost
- Convolutional neural networks

Data will be split into training/testing/validation sets and cross-validation will be used where appropriate. As there is variability by country and site country will be taken into account when splitting the data. Different methods for interpreting feature importance (which may involve different metrics for each type of model) and assessing the predictive performance (e.g. area under the ROC curve [AUC], Brier score) of models will be used.

5. Extension on the approaches in (4) to time-to-event outcomes, using death as the outcome.
6. Sensitivity analyses (such as restricting to lab-confirmed SARS-CoV-2, exploring variability by country by fitting models within country or groups of countries if feasible).

Handling of Missing Data
Analysis will be performed to ascertain a detailed overview of the extent of missingness in the data. This will enable us to identify the variables that are unsuitable for inclusion in the analysis. Depending on findings, different approaches for handling missing data will be considered and compared.

Other Information

Findings will be submitted for publication.

References

Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, Rumsfeld JS, Masoudi FA, Huang C, Normand SL, Mortazavi BJ, Krumholz HM. Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. JAMA Cardiol. 2021 Jun 1;6(6):633-641. doi: 10.1001/jamacardio.2021.0122. PMID: 33688915; PMCID: PMC7948114.

Sakr, Y., Giovini, M., Leone, M. *et al.* Pulmonary embolism in patients with coronavirus disease-2019 (COVID-19) pneumonia: a narrative review. *Ann. Intensive Care* 10, 124 (2020). https://doi.org/10.1186/s13613-020-00741-0