ISARIC

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

1.1 https://orcid.org/0000-0002-4471-7228

Analysis Plan for ISARIC International COVID-19 Patients

Title of proposed research
Development and Validation of a Prediction Tool for mortality and clinical deterioration of hospitalised COVID-19 patients
Version: (Date: Day/Month/Year)
11/04/2024
Working Group Chair (name, ORCID ID, email, institution, country)
Ibrahim Abubakar, 0000-0002-4471-7228, i.abubakar@ucl.ac.uk Institute of Global Health, University College London, UK.
[1]Working group co-chair (name, ORCID ID, email, institution, country)
Andrea Gori, 0000-0001-6587-4794, andrea.gori@unimi.it Department of Infectious Disease, Biomedical and Clinical Sciences, Ospedale Luigi Sacco, Milan, Italy
Statistician (name, ORCID ID, email, institution, country)
Dr Petr Andriushchenko, 0000-0002-4518-6588, petr.andriushchenko@iwr.uni-heidelberg.de, Heidelberg Institute of Global Health, Heidelberg University Hospital, Im Neuenheimer Feld 130.3, R. 313, 69120, Heidelberg, Germany.

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

[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.

be made publicly available on the ISARIC website. Please complete all sections of this form fully and return to ncov@isaric.org

Introduction

This study is part of the European Consortium, END VOC, led by Professor Ibrahim Abubakar at University College London (UCL). Integrating data from 19 partners and aimed at supporting the global response to the COVID-19 pandemic and variants of concern (VOCs) by integrating well-characterized cohorts and linking with existing European and international initiatives addressing 5 key aspects: 1) Detect and characterize emerging variants; 2) immune evasion and reinfections; 3) evaluating treatment escape segregated by VOC's; 4) investigating Long-COVID causes; 5) providing recommendations for management of outbreaks.

Fondazione IRCCS Ca' Granda Ospedale Maggiore, Policlinico in Milan, leads work package (WP)6, headed by Professor Andrea Gori. We focus on optimizing the clinical treatment and overall care of COVID-19 patients with emphasis on those infected with VOCs. One objective of ours is to produce a risk prediction tool that assesses the risk of disease deterioration in hospitalized COVID-19 patients with emphasis on treatment and VOCs. More details on the project can be found here: https://endvoc.eu/.

We plan a comprehensive analysis of COVID-19 patients collected from international cohorts. Centered on hospitalized COVID-19 patient data collected from January 2020 to December 2023, this diverse dataset would support thorough and elaborate analysis. Data sources include the END VOC database, specifically, the Italian COVID-19 Network. Efforts are also underway to gain access to the CH_SUR Hospital surveillance database in Switzerland and the national CVC-COVID-UK database (HES), as well as the ISARIC database upon approval of this application. This combination of data will offer a unique insight into the global experience of this pandemic.

Recognizing the shortcomings of existing models for predicting clinical deterioration and mortality in hospitalized COVID-19 patients, the lack of reliable clinical decision-making support, while armed with vast amounts of multicentric data from diverse geographical regions, we aim to produce a robust prediction framework leveraging machine learning (ML) techniques. The influx of ML risk stratification tools prompted by the pandemic are remarkably littered with flaws with many exhibiting high risk of bias, monocentric cohorts, limited sample size, or imprecision when adhering to reporting standards [1], [2]. According to a systematic review in October 2023 [1], only your very own ISARIC4C Deterioration model [3] was deemed suitable for clinical use, emphasizing the scarcity of validated models. Notably, data on COVID-19 VOCs were often neglected despite the literature demonstrating their association with disease progression [4], [5].

With the acute phase over, prioritizing disease management, particularly among high-risk groups has become crucial. Evaluation must be given to the impact of different treatments on patient outcomes, particularly in the context of emergent VOCs. To help assess this, our prediction tool will incorporate treatment data and variant information to predict in-hospital mortality and clinical deterioration accurately. This serves to enhance disease management, guide clinical decision-making, identify those at most risk, and optimize the allocation of resources [6].

Aligned with END VOC Consortium objectives, this study aims to bridge the gaps of current prediction tools through the integration of treatment strategies, VOCs information and traditional patient data providing an exhaustive risk stratification for COVID-19 mortality and clinical deterioration, thus facilitating informed decisions, optimized resource allocation, and refining disease management protocols. Data will be securely processed and harmonized via the UCL Trusted Research Environment (TRE) Data Safe Haven (DSH), and statistical analysis conducted by the Heidelberg Institute of Global Health (HIGH) and the Universitätsklinikum Heidelberg (UKHD). Several ML algorithms including logistic regression and ensemble-based models will be trained and rigorously evaluated through standardized validation procedures to optimize the algorithm before validating it on new datasets.

Participatory Approach

Our intention is to accumulate different cohorts from various countries via the END VOC project focused on COVID-19 VOCs in relation to transmission, disease treatment, prevention and pandemic prediction. To achieve this, our study will be conducted according to the END VOC Policies including the declaration of Helsinki. Data collection and storage will be compliant with ICH-GCP requirements and the EU & UK General Data Protection Regulation (GDPR).

A data sharing agreement is signed by all partners involved. Access to data will be applied for, harmonized and stored on the TRE DSH hosted by UCL. Analyses will occur on the UCL server; therefore, all datasets will be protected in accordance with END VOC policy. All necessary ethical approval will be obtained and declared within the project.

We are committed to acknowledging all contributors. Adhering to ethical publishing standards, we ensure those who substantially contribute to conception, design, acquisition, analysis, or interpretation of data are appropriately recognized as authors. Contributors who do not meet the criteria for authorship but provide valuable assistance will be acknowledged in the appropriate section.

Upon completion, we aim to submit our findings to peer-reviewed scientific journals specializing in artificial intelligence in healthcare, infectious diseases, and epidemiology. Publication timelines depend on the analysis complexity, the extent of peer review, and journal editorial processes. Submission is anticipated within 6 to 12 months after data collection.

Additionally, we plan to disseminate our research through conferences, seminars, and workshops attended by healthcare professionals, researchers, policymakers, and stakeholders in the field of infectious diseases and public health. Presenting at national and international conferences to reach a broad audience and foster discussions on the application of ML in predicting clinical outcomes for COVID-19 patients.

Research Plan

Summary of Research Objectives

This research seeks to address the limitations of current prediction tools for clinical outcomes for COVID-19 patients. Employing the latest ML techniques and geographically diverse data, the objective is to develop and validate a reliable prediction tool to predict in-hospital mortality and clinical deterioration of COVID-19 hospitalized patients, integrating treatment information and VOCs data information along with traditional patient data. Clinical deterioration is defined as admission or transfer to ICU, ventilatory support (non-invasive, invasive mechanical, or extracorporeal membrane ventilation), and length of stay in ICU.

The tool will be provided on a user-friendly, open-source platform, updating accordingly to incorporate new treatments and VOC data.

We hypothesize that a prediction tool with the integration of treatment and variant data alongside conventional predictors will supply a dependable, and well-grounded prediction tool for clinical outcomes of COVID-19 patients with the potential to serve as a valuable framework for other diseases.

Proposed Target Population

Hospitalized COVID-19 patients.

Clinical Questions/Descriptive Analyses

1. What are the most significant predictors of clinical deterioration and in-hospital mortality among COVID-19 patients when considering demographic, clinical, laboratory, treatment, and VOC data?

2. How does incorporation of variant data impact the sensitivity and specificity of the prediction model for calculating the risk of disease severity or treatment response variability in COVID-19 patients?

3. What specific treatment variables have the most significant impact on predictive accuracy and how do they interact with other predictors?

Planned Statistical Analyses, Methodology and Representation

Please note, data has not yet been accessed. Therefore, these methods reflect anticipated data. A detailed data analysis protocol will be developed once data are obtained, and a preliminary analysis conducted

Study Population

Data of hospitalised COVID-19 patients will be extracted spanning January 2020 to December 2023. Databases include cohorts from END VOC, specifically the COVID-19 Network. Efforts are underway for access to the CH_SUR hospital surveillance database and the national UK database, CVC-COVID-UK (HES). We are continuously exploring for potential integration of supplementary cohorts in the analysis.

Inclusion criteria consist of adults (≥18 years) with SARS-CoV-2 infection. Confirmation of infection via RT-PCR, lateral flow, nasal swabs, or rapid diagnostic tests (RDTs). No exclusion criteria.

The study will adhere to the TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) guidelines [7], TRIPOD – Cluster checklist [8] and the PROBLAST protocol to ensure a low risk of bias [9].

Data from external databases will be securely migrated to the TRE DSH warehouse hosted by UCL. Model development will be spearheaded by HIGH.

Data Pre-processing

Extracted data will undergo harmonization and (if not already) anonymization into a unified dataset within the TRE DSH data warehouse. Secure processing in the TRE DSH will involve the unification of features and values, data cleaning ensuring consistent values and properly converted to required data types, encoding categorical variables, standardising, or normalising numerical features to bring them to a uniform scale.

Normal data parameters will be defined, and outliers identified with statistical methods. Severely corrupted data missing a substantial number of values, i.e., more than 10%, will be omitted.

Feature Selection

Identification of the most significant features prevents overfitting, using correlation statistics such as chi-square for categorical variables, Pearson correlation coefficient for continuous variables, mutual information etc. Features showing high correlation with others will be discarded. Additionally, the dataset will be checked for any imbalance.

Model Development and Hyperparameter Optimization

Stratified sampling will divide the harmonized dataset into development (training) and test sets. If granted access, we intend to use the national UK database, CVC-COVID-UK (HES) for the validation dataset. Several ML models will be trained on the development dataset and assessed by performance on the validation dataset. The best models will then be evaluated on the test sample.

We intend to assess traditional logistic regression approaches with lasso, Ridge and elastic net (Lasso and Ridge combined) regularizations, alongside some ensemble-based models like XGBoost.

Evaluation metrics include C statistics (AUC), 95% confidence interval, sensitivity, specificity, and predictive values of the models to comprehensively analyse model performance. As per the TRIPOD guidelines, the clinical applicability of the tool will be assessed through calculation of the net benefit [10]:

$$\textbf{True positive/n - True negative/n} \times \textbf{(Pt/ (1 – Pt)}$$

Key Variables and Definitions

Demographic: Age, gender, ethnicity, body mass index.

Clinical: Data of admission, data of primary outcome or discharge, onset symptoms, comorbidities, vital signs at admission, disease severity, oxygen required during hospitalization, admission to intensive care unit (ICU), and ventilation support.

Laboratory exams [on day of admission]: total blood count, inflammatory biomarkers and organ damage biomarkers.

Treatment*:* anti-COVID-19 treatment administered before and during the hospitalisation, dosage administered, dates of administration, date of treatment termination.

Variant/VOC: variant/sub-lineage responsible for the infection. If missing information, we intend to discriminate VOCs based on the national SARS-CoV-2 genetic diversity monitoring in Switzerland (link), and in the UK (link), only considering intervals where specific VOCs were dominant.

Potential Confounders

Potential confounders, such as insufficient data, data quality, outliers and missing values will be managed through randomization. Resampling methods such as cross-validation will be used to randomize the training and validation sets, bootstrapping can be used for handling unbalanced sets and Multiple Imputation for the imputation of missing variables[11]. The training and test datasets will be randomly stratified to give large heterogeneity, and validation of the prediction models using large independent datasets will ensure the generalizability of our findings across different patient populations and healthcare settings.

The severity of COVID-19 upon hospital admission can vary widely among patients and may influence both treatment decisions and clinical outcomes. Hence, we will use multivariable regression models to adjust for known confounders, assess their independent associations with clinical outcomes and perform sensitivity analyses to evaluate the robustness of our findings to different model approaches and assumptions about potential confounders [12].

Effect Modifiers

Variations in treatment protocols and therapies for COVID-19 may modify the relationship between predictor variables and clinical outcomes. Moreover, different VOCs may exhibit varying levels of virulence and response to treatment, while vaccination status of patient potential alters the risk of clinical deterioration and mortality. We will conduct subgroup analyses to explore potential effect modification by these key variables and perform sensitivity analyses.

Handling of Missing Data

Preliminary analysis would be performed to ascertain a detailed overview of the extent of missingness in the data. This should enable the identification of variables which lack sufficient data to allow for any useful analysis to be performed on them. Type of missingness shall be considered including whether data are not missing at random and follow-up with sites will be conducted if appropriate. Variables with greater than 30% missingness will be excluded from the analysis. Where appropriate, imputation will be performed using k-nearest Neighbour (kNN) and iterative imputation methods will estimate and impute the values.

Other Information

References

[1] K. S. Appel, R. Geisler, D. Maier, O. Miljukov, S. M. Hopff, and J. J. Vehreschild, “A Systematic Review of Predictor Composition, Outcomes, Risk of Bias, and Validation of Coronavirus Disease 2019 (COVID-19) Prognostic Scores,” *Clinical Infectious Diseases*, Oct. 2023, doi: 10.1093/cid/ciad618.

[2] C. Buttia *et al.*, “Prognostic models in COVID-19 infection that predict severity: a systematic review,” *European Journal of Epidemiology*, vol. 38, no. 4. Springer Science and Business Media B.V., pp. 355–372, Apr. 01, 2023. doi: 10.1007/s10654-023-00973-x.

[3] R. K. Gupta *et al.*, “Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study,” *Lancet Respir Med*, vol. 9, no. 4, pp. 349–359, Apr. 2021, doi: 10.1016/S2213-2600(20)30559-2.

[4] R. Levi, E. G. Zerhouni, and S. Altuvia, “Predicting the spread of SARS-CoV-2 variants: An artificial intelligence enabled early detection,” *PNAS Nexus*, vol. 3, no. 1, Dec. 2023, doi: 10.1093/pnasnexus/pgad424.

[5] R. Lorenzo-Redondo, A. M. de Sant’Anna Carvalho, J. F. Hultquist, and E. A. Ozer, “SARS-CoV-2 genomics and impact on clinical care for COVID-19,” *Journal of Antimicrobial Chemotherapy*, vol. 78, pp. II25–II36, Nov. 2023, doi: 10.1093/jac/dkad309.

[6] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, “Machine learning in clinical decision making,” *Med*, vol. 2, no. 6. Cell Press, pp. 642–665, Jun. 11, 2021. doi: 10.1016/j.medj.2021.04.006.

[7] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement,” *Ann Intern Med*, vol. 162, no. 1, pp. 55–63, Jan. 2015, doi: 10.7326/M14-0697.

[8] T. P. A. Debray *et al.*, "Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist," *BMJ*, 2023, doi: 10.1136/bmj-2022-071018.

[9] R. F. Wolff *et al.*, "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," *Ann Intern Med*, vol. 170, no. 1, pp. 51–58, Jan. 2019, doi: 10.7326/M18-1376.

[10] D. Piovani, R. Sokou, A. G. Tsantes, A. S. Vitello, and S. Bonovas, "Optimizing Clinical Decision Making with Decision Curve Analysis: Insights for Clinical Investigators," *Healthcare (Switzerland)*, vol. 11, no. 16, Aug. 2023, doi: 10.3390/healthcare11162244.

[11] M. K. Nariya, C. E. Mills, P. K. Sorger, and A. Sokolov, "Paired evaluation of machine-learning models characterizes effects of confounders and outliers," *Patterns*, vol. 4, no. 8, Aug. 2023, doi: 10.1016/j.patter.2023.100791.

[12] G. Lippi, C. Mattiuzzi, and B. M. Henry, "Uncontrolled confounding in COVID-19 epidemiology," *Diagnosis*, vol. 10, no. 2. Walter de Gruyter GmbH, pp. 200–202, May 01, 2023. doi: 10.1515/dx-2022-0128.