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

Please complete the following sections:

Title of proposed research
Predicting Long-COVID Syndrome Machine Learning Models (PRELOCS)
Version: (Date: Day/Month/Year)
Version 1: 25/02/2022 Version 2: 4/05/2022
Working Group Chair (name, ORCID ID, email, institution, country)
Dr Kalaiarasu M. Peariasamy, ISARIC, Institute for Clinical Research, National Institutes of Health Malaysia Email: drkalaiarasu@gmail.com ORCID ID: 0000-0001-9279-3498
[1]Working group co-chair (name, ORCID ID, email, institution, country)
Dr Pei Xuan Kuan, Institute for Clinical Research, National Institutes of Health Malaysia Email: pxkuan@gmail.com ORCID ID: 0000-0001-7041-571X Dr Mohan Dass Pathmanathan, Institute for Clinical Research, National Institutes of Health Malaysia Email: mohan.pathmanathan@gmail.com

[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.

ORCID ID: 0000-0001-8385-0315

Kian Boon Law, Institute for Clinical Research, National Institutes of Health Malaysia
Email: kblaw@crc.gov.my
ORCID ID: 0000-0002-1175-1307
Dr Xin Ci Wong, Institute for Clinical Research, National Institutes of Health Malaysia
Email: wongcx.crc@gmail.com
ORCID ID: 0000-0002-1036-8023

Dr Mohd Aizuddin Abdul Rahman, Institute for Clinical Research, National Institutes of Health Malaysia
Email: drmohdaizu@gmail.com
ORCID ID: 0000-0002-0014-1060

Prof Dr Nai Ming Lai, Taylor's University Malaysia
email: NaiMing.Lai@taylors.edu.my
ORCID ID: 0000-0002-4204-3098

Statistician (name, ORCID ID, email, institution, country)

Kian Boon Law, Institute for Clinical Research, National Institutes of Health Malaysia
Email: kblaw@crc.gov.my
ORCID ID: 0000-0002-1175-1307

Prof Dr Kok Wei Khong, Taylor's University Malaysia
email: kokwei.khong@taylors.edu.my
ORCID ID: 0000-0003-2603-1545

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

This document details the initial analysis plan for publication. The analysis builds on data collected from long-COVID patients using the dataset from ISARIC database.

Coronavirus disease 2019 (COVID-19) may cause prolonged symptoms to some patients lasting up to months affecting the health and quality of life of patients. This post-COVID syndrome is defined as long-COVID by the World Health Organization (WHO). Similar terms include long-haul COVID, long-term effects of COVID, chronic COVID or post-COVID-19 syndrome (1, 2). Long-COVID patients

often present with a spectrum of prolonged symptoms, including mainly fatigue, shortness of breath, cough, joint pain, and chest pain. Other non-specific symptoms include cognitive impairments, headache or dizziness, insomnia, depression, anxiety, loss of appetite, diarrhoea, intermittent fever, anosmia, ageusia, sore throat and rashes (1,2). The long-COVID patients are categorized under a disability group based on the Americans with Disabilities Act (ADA) (2).

The management of long-COVID syndrome is mainly supportive. However, rehabilitation services are required for patients with profound symptoms that have resulted in substantial burden and disability (1). Currently, mechanisms, associated factors and possible interventions for long-COVID are being studied intensively across the world, which aims to improve outcomes and reduce global burden that is associated with this condition (2).

Machine learning (ML), a subset of artificial intelligence (AI) has been widely adopted in contact tracing, cluster analysis, screening, diagnosis, and prognostication of COVID-19 disease during this pandemic. Through the advancement in computing technologies, huge volumes of data could be processed and analyzed to identify specific patterns and intuitions to forecast the spread of COVID-19 (3). Machine learning has also been used in certain applications such as improving and supporting remote communication, providing knowledge on the spread of COVID-19, and producing rapid indicators for predicting trends with guidance for interventions (3). Machine-learning enabled chatbots have been used in the healthcare settings as well as government institutions for contactless screening of COVID-19 and answering general queries from the public (3).

However, to-date, research is still lacking on the use of ML to generate a predictive model for post-COVID syndrome. Thus, we aim to develop a ML model for early detection of patients who are at risk of long-COVID syndromes for early initiation of medical treatment to prevent further complications.

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 in the cited authorship 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

The primary objective of this study is to develop, test and validate machine learning models for the early detection of patients who are at risk of long-COVID syndromes at Asia region.

Proposed Target Population

All Asia region data involving patients diagnosed with long -COVID syndrome and undergoing follow-up.

Inclusion Criteria:
All patients diagnosed and recovered from COVID-19 infection with follow-up with long-COVID syndrome. We follow the latest WHO case definition of long-COVID, as follows: "Post COVID-19 condition occurs in individuals with a history of probable or confirmed SARS-CoV-2 infection, usually three months from the onset of COVID-19 with symptoms that last for at least two months and cannot be explained by an alternative diagnosis. Common symptoms include fatigue, shortness of breath, cognitive dysfunction but also others and generally have an impact on everyday functioning. Symptoms may be new onset following initial recovery from an acute COVID-19 episode or persist from the initial illness. Symptoms may also fluctuate or relapse over time." (4)

We will adapt the definition where appropriate in light of new evidence.

Exclusion Criteria:

1. Suspected cases without laboratory or health authorities or test kit confirmation COVID-19.

Clinical Questions/Descriptive Analyses

1. What are the significant predictors for early identification of patients with risk of developing long-COVID syndrome?
2. What is the best machine learning algorithm model for early detection of long-COVID syndrome in Malaysia setting?

Planned Statistical Analyses, Methodology and Representation

This a cross-sectional study using available data sources from ISARIC database and Malaysia's ongoing COVID-19 study for the development, testing and validation of machine learning models. The study will be conducted in two phases as described below:

Phase I: Development, training, testing and validation of machine learning models
Data extraction and data cleaning will be performed by the study team. Data will be used to identify relevant predictors for building various machine learning models for training and testing. The extracted dataset will first be studied to identify relevant predictors using univariable and multivariable logistic regression methods. Odds ratio (OR) and 95% confidence intervals (95% CI), test statistics and p-values will be compiled and compared to assess the predicting potential and weight of all parameters included. Following the selection of potential parameters, the dataset

will be divided into training, testing and validation datasets in applicable ratio. Machine learning models studied include neural network (NN), support vector machine (SVM), random forest (NF), k-neighbor nearest (kNN), logistics regression (LR) and so on will be trained and tested to identify the best ML model for further validation.

Phase II: Deployment of machine learning models

In phase two of the study, we will further validate the accuracy and utility of the ML model against real-world data from the on-going COVID-related clinical data from the on-going clinical dataset of the current and newly discharged patients managed by the Malaysian Ministry of Health hospitals. Patients who are currently under follow-up will be monitored according to their risk of developing long-COVID syndrome as predicted by the ML model, and the actual outcomes in terms of the occurrence of long-COVID syndrome evaluated against the predicted likelihood of occurrence. We will pilot the use of the ML model in the real-world with patients from selected local tertiary hospitals before wider deployment. To enable end-user accessibility and robust support and troubleshooting, front end dashboard will be developed, and back-end resources incorporated to the established ML model towards a final user-ready version.

This model could be used to identify post COVID patients at risk of long COVID for future trials of preventive interventional study and post COVID management.

In order to evaluate our methodology, R package software and SAS Enterprise Miner/Guide software will be used to automate predictive modeling in logistic regression (LR).

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. Type of missingness shall be considered including whether data are not missing at random. Variables with greater than 30% missingness will be excluded from analysis. Where appropriate, imputation will be performed using Multiple Imputation by Chained Equations (MICE). Complete case analysis will be performed as sensitivity analysis.

Other Information

The timeline of study from Phase 1 to completion of Phase 2 will take approximately 18 months. The aim is to submit the analysis for publication by the 1st quarter of 2024. The results will be disseminated via peer-reviewed publication and translation of the findings into national policies as appropriate.

References

1. Long-term effects of coronavirus (long COVID). National Health Service (NHS). United Kingdom. [internet]. 2021. Available from:

https://www.nhs.uk/conditions/coronavirus-covid-19/long-term-effects-of-coronavirus-long-covid/
2. Post-COVID conditions. Centers for disease control and prevention (CDC). [internet]. 2021. Available from: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html
3. How AI and machine learning are helping to fight COVID-19. World Economic Forum. [internet]. 2021. Available from: https://www.weforum.org/agenda/2020/05/how-ai-and-machine-learning-are-helping-to-fight-covid-19/
4. WHO. A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021. WHO publications [internet]. 2021. Available from: WHO/2019-nCoV/Post COVID-19 condition/Clinical_case_definition/2021.1
5. Syeda HB, Syed M, Sexton KW, et al. Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review. *JMIR Med Inform*. 2021;9(1):e23811. Published 2021 Jan 11. Doi:10.2196/23811
6. A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021. World Health Organization. [internet]. 2021. https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1
7. COVID-19 rapid guideline: managing the long-term effects of COVID-19. NICE Guideline 2020.https://www.nice.org.uk/guidance/ng188/resources/covid19-rapid-guideline-managing-the-longterm-effects-of-covid19-pdf-66142028400325
8. Taquet M, Dercon Q, Luciano S, Geddes JR, Husain M, Harrison PJ (2021) Incidence, co- occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19. PloS Med 18(9): e1003773. https://doi.org/10.1371/journal. Pmed.1003773