Analysis Plan for ISARIC International COVID-19 Critically Ill Cohort

September 2020

Introduction

Scope of document

This document details the initial analysis plan for publication on a subset of critically ill individuals in the global cohort in the ISARIC database, as of 03 Aug 2020. There are currently 44 countries (as of 3 AUG 2020) contributing data and these have so far contributed data on 102,038 patients. This data will represent the global experience of the first 6 months of this pandemic.

Rationale for project

A descriptive dataset of critically ill patients around the world with COVID-19 has been identified in ISARIC global collaborator meetings as a priority outcome. There are few international cohorts of this critically ill population described in the literature and so this collaborative project hopes to address this. This collaboration aims to describe the first and subsequent waves of the COVID-19 pandemic.

Project aims

1) To summarize the demographic characteristics and clinical features of ~16,000 critically ill patients of any age, admitted to hospital with COVID-19 across high-income, middle-income, and low-income settings, across temporal phases of the pandemic.
2) To describe clinical outcomes (e.g., ICU length of stay, hospital length of stay, ICU mortality, hospital mortality, days of mechanical ventilation, among others) of critically ill patients with severe COVID-19.
3) To characterize the variability in the clinical features and management strategies of these patients.
4) To explore the risk and protective factors associated with mortality for these patients.
5) To determine variation over time in various phases of ICU 'journey' including treatments received, length of stay in ICU and duration of mechanical ventilation.

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.

Data

Intended datasets for inclusion

Datasets from all sites are eligible for inclusion in this analysis.

Critically ill patients will be defined as those who meet a modified WHO definition of critical disease, ie those requiring ventilatory support (i.e., invasive/non-invasive mechanical ventilation and High Flow Nasal Cannula) and/or vasopressors. Sensitivity analyses will be performed for patients requiring an ICU/HDU admission, as self-defined.

Exclusion criteria at an individual patient level

1) For each analysis, individuals with incomplete data on the outcome of interest shall be excluded, except with regards to survival analysis where appropriate censoring will be applied to those individuals with incomplete data.

Research questions

Clinical Question	Planned statistical analyses	Planned representation in manuscript(s)
Univariable/descriptive analyses		
1) What are the characteristics of critically ill patients with respect to key demographic variables (e.g., age, gender, comorbidities, healthcare-workers, and ethnicity), continents and regional income/ human development index stratification. 2) What proportion of critically ill patients are: discharged alive, have died, have residual technological dependence, or are still in hospital at a censored time point (Day 14, 28 and 90)). Sub-divide populations by (i) all (ii) those ever-receiving invasive ventilation (iii) those ever-receiving non-invasive, invasive ventilation or high flow nasal oxygen. Then, stratify	1) Overall frequencies of key demographic variables as well as frequencies stratified by region. 2) Overall proportions of deaths, recoveries, and ongoing admissions. Stratify frequencies by region. 3) Prevalence of Comorbidities/Symptoms/Treatments (CSTs) stratified by age group, sex and confidence intervals (CIs) of prevalence estimates.	Bar plots – for displaying the frequencies of categorical variables Box plots – for summarizing distributions (quantitative outcome variables only) UpSet plots – for displaying frequencies of combinations of CSTs

the data by the moment of the pandemic the patient was admitted. 3) What are the symptoms, vital signs, and laboratory findings at time of admission, and does this differ with age and region (continent)? 4) What proportion of critically ill patients received different levels of supportive care (e.g. O2/HFNC/ NIV/IV/ECMO/RRT/prone positioning/vasopressors), and other treatments (e.g., antibiotics, antivirals, steroids, anticoagulation, among others). What proportion of those receiving (i) non-invasive ventilation or (ii) high flow nasal oxygen required eventual mechanical ventilation. 5) What is the distribution of key time variables (length of hospital admission, length of ICU admission, time from symptom onset to ICU admission, length of time requiring IMV/NIV/ECMO/high flow nasal cannula, time to discharge/death). Moreover, describe the outcomes and multiple ventilatory supportive interventions in the same patient. 6) What proportion of critically ill patients develop COVID-19 related complications (e.g., pulmonary embolism, myocardial infarctions,	4) Proportion of patients requiring IMV/NIV ECMO/high flow nasal cannula/ prone position. 5) Summaries (mean, median and SD) of key time variables. A Gamma model would be used to provide alternative summaries which account for censoring. 6) A Kaplan-Meier–based competing risk approach [2] would be employed in the estimation of the overall CFP as well as the CFP for the indicated subgroups. A Cox proportional-hazards model will be applied to estimate unadjusted and adjusted HR. 7) Summaries of vital signs and laboratory results stratified by age group. 8) Contingency tables of infectious complications	Summary tables Survival curves – for displaying cumulative probabilities for death/discharge

heart failure among others) during hospital admission.		
7) What is the overall case-fatality proportion (CFP) for this cohort? How have the probabilities of death and discharge varied over time? Does the CFP differ by age, sex, or region (continent or by regional income stratification, and time/epoch of the pandemic thus far?		
8) Are there differences in the distribution of vital signs and laboratory results at presentation at hospital and at ICU by age group and in patients requiring mechanical ventilation.		
9) Time spent on oxygen supplementation and mechanical ventilation (subdivided by modality)		
10) Does seasonal variation contribute to the density of ICU admissions for COVID-19 illness by geographic regions		
Bivariable analyses		
1) Do the distributions of key demographic variables, co-morbidities, management strategies and outcomes, differ by outcome and region [1]	1) Chi-square tests for the differences in age/sex/ethnicity distribution by age, region and outcome.	Correlogram - for displaying correlations

2) Do the distributions of key time variables differ by age, sex or region? Of particular interest is whether the duration of symptom onset to ICU admission varies by region, or the duration of ICU stay and duration of specific intervention (i.e., invasive ventilation, HFNC, NIV) varies by region and stratified by before/after invasive ventilation. 3) Are there temporal changes in the demographics and management (i.e., medications, ventilation types, use of ECMO) and outcomes of critically ill patients over the first six months of the pandemic?	2) Chi-square tests for the differences in outcome proportions by region. 3) Chi-square tests for the differences in CST proportions by age, region and outcome. 4) Chi-square tests for pairs of CSTs and phi correlation coefficient for significant comparisons. 5) One-way ANOVA for comparing samples of key time variables for age, sex, and region categories. 6) Survival Analysis 7) Time series analyses/discrete epoch analyses (either 3 weekly or monthly or bi-monthly) informed by the data	Violin plots – for visualizing the distribution of quantitative variables (e.g. time to discharge/death)
Multivariable analyses		

1) What clinical and laboratory factors at ICU admission predict death in critically ill patients with COVID-19? – Develop the ISARIC COVID Score. 2) COVID-19 phenotypes. 3) Report the association of conventional clinical interventions/treatments, including prone ventilation, corticosteroids, remdesivir, tocilizumab, among others, on clinical outcomes. 4) Survival analysis by clinical characteristics and other predictors.	Logistic regression or Cox regression.[2]. Cluster analysis Latent Class Analysis Marginal structural modelling Machine Learning ensemble method Predictor variables. ● Demographic: (age, gender, comorbidities and ethnicity) ● Clinical characteristics (symptoms, vital signs, a laboratory values) ● Requirement for ventilation ● Vital signs at presentation ● Laboratory findings at presentation ● Use of RRT. ● Use of ECMO. Outcome variables ● Mortality	Alluvial plots Chord diagrams Dendrograms t-SNE plots Cumulative incidence function plots

[1] Countries will be aggregated by WHO or World Bank region. To ensure that we have large enough sample sizes to detect effects where present, regions with less than 100 patients may be excluded from comparison analyses. This is to also to ensure a fair representation of the various outcomes or variables of interest across regions to be compared.

[2] The method employed would depend on the level of censoring present in the data.

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 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 analysis. Where appropriate, imputation will be performed using Multiple Imputation by Chained Equations (MICE).

Statistical procedures

Summary statistics including contingency tables, median/IQR, and mean/SD will be used for appropriate data type of distribution of each variable. Association between categorical variables will be assessed using Chi-squared tests, association between categorical dependent variables and continuous outcomes will be assessed using analysis of variance or t-tests dependent on number of categories. Association between two continuous variables will be assessed using Pearson's and Spearman's correlation.

Prediction modelling of events will proceed by initially selecting all potential predictors and confounders. Known confounders will be included regardless of their association with the outcome. A k-folds cross-validation scheme will be implemented where the folds will be based on geographic region, which should provide better approximation to the out-of-sample error. Co-linearity and association to outcome of predictors will be assessed using scatter matrices and correlation plots. Distribution of predictors will be examined and transformed as appropriate. Variable selection and model fitting shall be incorporated into a single design using an adaptive elastic-net logistic regression model. Model fit will be assessed using Akaike and Bayesian Information Criterion, area under the receiver operator characteristic curve, and other binary classifier accuracy measures. Validation will be assessed using the same metrics on the predicted probabilities of the test set.

In time-to-event analysis care will be taken to ensure all assumptions of statistical models are met including proportional hazard assumption, and presence of immortal time. These will be assessed graphically using Schoenfeld residuals compared to each covariate and Kaplan-Meier survival probability plots. Parametric models will be considered if proportional hazards assumptions are violated. Frailty models will be used to account for random effects between regions. The cumulative incidence function will be used for comparison between groups, and proportional sub-distribution hazard (Fine & Gray) models will be used in regression analysis.

For COVID-19 phenotyping, appropriate clinical indicators will be identified and assessed for their appropriateness using their distribution, missingness, and correlation with main outcomes. The data shall be initially split 4:1 into a training and test set. Clustering of indicators will be assessed through a series of OPTICS and DBSCAN plots. Determination of optimal number of clusters will be performed by examining phenotype size, consensus matrix heatmaps, consensus cumulative distribution function plots, and by examining the change in consensus CDF AUC for cluster number. The resulting clusters will be analyzed using t-SNE, and hierarchical clustering dendrograms. Prediction of clusters on test set will

be applied and measures of clustering including Calinski-Harabasz index, Davies-Bouldin index, and silhouette coefficient will be used to validate phenotypes.

Sensitivity analyses

Robustness of clinical phenotypes will be assessed by re-estimating and predicting clusters through multiple random splits of data. Influence of each variable on identified clusters will be assessed by leaving out each variable and re-fitting and examining predicted clusters with original fit through a series of clustering metrics including V-measure and homogeneity score.

The unmatched logistic regression model will be compared to a matched design where propensity score matching occurs on known confounders and the resulting odds ratios are compared to the unmatched design.

Should imputation be applied to variables, a complete case analysis will also be performed.

Sensitivity to definition of ICU admission will be explored by comparing resulting analyses to analyses where only those individuals admitted to a high-dependency unit, non-invasive ventilation or high-flow nasal cannula, intravenous vasoactive medications, or invasive ventilation are excluded.

Survival analyses that incorporate incomplete data on individual outcomes will be compared to a complete data analysis where all incomplete cases are excluded.

Subsequent analyses

Plans for subsequent analyses to include other specific analyses as they emerge, including descriptive analyses of ventilation patterns in intubated patients, descriptive analyses of patients who die without receiving ICU care or ventilation, and using machine learning models to predict outcomes based on presenting characteristics.

Manuscripts

1. **Description of critically ill patients included in the total dataset:** Including a comparison of outcomes based on patients admitted to the ICU and those meeting the WHO definition for being critically ill. This will include the stratified analyses, by epoch during the pandemic, that the patient was admitted and the geographical location/income, and identification of risk factors in critically ill patients.
2. **Identification of the phenotypes of critically ill patients:** We will report phenotypes of patients with severe COVID-19 and determine the differences in clinical outcomes among each phenotype.
3. **Ventilatory failure in patients with COVID-19:** We will describe the patterns of ventilatory support in patients who require it for COVID-19, including ECMO (HFNC, NIV, IMV, ECMO). This will include analyses of geographic/regional variation in support, including variations in time.
4. **COVID-19 related complications and the effect on outcomes.** A fundamental clinical question is about major adverse cardiovascular events (MACE) and pulmonary emboli in patients with COVID-19. Thus, here we should describe the frequencies of these complications, determine the impact of these complications in clinical outcomes, and determine the risk factors.

Software

All analyses will be performed in R [1].

SPSS 27 for MAC.

References

1. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
2. Ghani, A. C., Donnelly, C. A., Cox, D. R., Griffin, J. T., Fraser, C., Lam, T. H., ... & Leung, G. M. (2005). Methods for estimating the case fatality ratio for a novel, emerging infectious disease. American Journal of Epidemiology, 162(5), 479-486