Analysis Plan for ISARIC International COVID-19 Cohort

Overall descriptive analysis

July 2020

Introduction

Scope of document

This document details the analysis plan for publication on the non-UK cohort in the ISARIC database. There are currently 36 countries (as of 8th June 2020, excluding UK data which has been reported elsewhere) contributing data and these have so far contributed data on 10941 patients.

This proposed plan includes all analyses in the latest ISARIC report (8th June) and is mainly motivated by the research questions raised by ISARIC partners as outlined in the next section.

Rationale for project

An overall descriptive dataset has been identified in collaborator meetings as a priority outcome. There are few international cohorts described in the literature and so this collaborative project hopes to address this.

Project aims

1) To summarize the demographic characteristics and clinical features of 10941 patients admitted to hospital with COVID-19 across high-income, middle-income, and low-income settings.
2) To characterise the variability in the clinical features of these patients.

3) To explore the risk factors associated with mortality and ICU admission for these patients.

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.

Data

Intended datasets for inclusion

Datasets from all sites outside of the UK are eligible for inclusion in this analysis.

Exclusion criteria at an individual patient level

1) Patients who have
    i) a negative laboratory result for SARS-CoV-2 , or
    ii) an enrolment date less than 14 days[1] before the analysis reference date

    shall be excluded from the study.

2) For each analysis, individuals with incomplete data on the variables of interest shall be excluded.[2]

[1] By excluding patients enrolled less than two weeks before the reference date, we aim to reduce the number of incomplete data records and thus improve the generalisability of the results and the accuracy of the outcomes. The 14-day threshold is subject to change however.

[2] To limit the rate of missingness, the ISARIC team has set up a comprehensive protocol for coordinating with sites on all matters related to data completeness.

Research questions

Clinical Question	Planned statistical analyses	Planned representation in manuscript(s)
Univariable/descriptive analyses		
1) What are the characteristics of patients with respect to key demographic variables (age, sex and ethnicity) and region? 2) What proportion of patients are: discharged alive, have died and still in hospital? 3) How often do specific comorbidities/symptoms/treatments (CSTs) occur? 4) What proportion of patients require different levels of supportive care (e.g. O2)? 5) What is the distribution of key time variables (length of hospital admission, length of ICU admission, time from symptom onset to admission, length of time requiring IMV/NIV/ECMO/high flow nasal cannula, time to discharge/death)? 6) What is the overall case-fatality ratio (CFR) for this cohort? How have the probabilities of death and discharge varied over time? Does the CFR differ by age, sex or region? 7) Are there differences in the distribution of vital signs and laboratory results at presentation by age group?	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 CSTs stratified by age group and confidence intervals (CIs) of prevalence estimates. 4) Proportion of patients requiring IMV/NIV ECMO/high flow nasal cannula. 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 nonparametric Kaplan-Meier–based competing risk approach [2] would be employed in the estimation of the overall CFR as well as the CFR for the indicated subgroups. 7) Summaries of vital signs and laboratory results stratified by age group.	• 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 • Summary tables • Survival curves – for displaying cumulative probabilities for death/discharge

Bivariable analyses		
1) Do the distributions of key demographic variables differ by outcome and region[3]? 2) Does the distribution of outcomes differ by region? 3) Are there differences in the prevalence of CSTs by age group and region? 4) Which combinations of CSTs co-occur? 5) Do the distributions of key time variables differ by age, sex or region? Of particular interest is whether the duration of symptom onset to presentation varies by region (this may help us better characterize health care seeking behavior across various geographical and income settings).	1) Chi-square tests for the differences in age/sex/ethnicity distribution by age, region and outcome. 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.	• Correlogram - for displaying correlations • Violin plots – for visualizing the distribution of quantitative variables (e.g. time to discharge/death)
Multivariable analyses		
1. What clinical and laboratory factors predict poor outcome in hospitalized patients with COVID-19?	Logistic regression or Cox regression[4]. <u>Predictor variables</u>. • Demographic : (age, gender, and ethnicity) • Clinical (Top 5 most frequently occurring CSTs) • Vital signs at presentation • Laboratory findings at presentation <u>Outcome variables</u>	• Forest Plot

[3] 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.
[4] The method employed would depend on the level of censoring present in the data.

	• Mortality • IMV/NIV/ECMO requirement[5]	

[5] The analysis may be run on different patient subgroups, based on the number of days spent in hospital prior to IMV/NIV requirement. Patients who began IMV/NIV treatment less than 3 days after hospital admission would not be considered. The threshold of 3 is subject to an upward revision.

Statistical Considerations

1. 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. Aside follow-up with sites, missing data may be handled by employing multiple imputation techniques. Actions taken would depend on the type of missingness and the relevance of the variable to the analysis under consideration.

2. Assumptions underlying all statistical tests would be checked prior to use. Where assumptions are not met, alternative non-parametric techniques would be explored. Otherwise, the analysis under consideration would not be pursued.

Software

All analyses will be performed in R [1].

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.