
# Documentation

Technical Documentation

*Ahmed Alaa, Zhaozhi Qian, and Mihaela van der Schaar*

April 14, 2020

**Title:** Documentation
**Version:** 0.1
**Issued by:** Cambridge Center for AI in Medicine
**Creation date:** April 14, 2020

**File history:**

| | |
|---|---|
| **Date** | April 14, 2020 |
| **Author** | Ahmed Alaa, Zhaozhi Qian, and Mihaela van der Schaar |
| **Description** | Documentation for the Adjutorium COVID-19 model |

# Contents

# 1 Data Cleaning and Processing

The COVID-19 Hospitalisation in England Surveillance System (CHESS) collects epidemiological data (demographics, risk factors, clinical information on severity, and outcome) on COVID-19 infection in persons requiring hospitalisation and ICU. This section provides a description of the data, explains our data processing steps, along with some basic statistics of the cleaned data set.

## 1.1 CHESS data description

### 1.1.1 Data file contents

The data residing in file "CHESS_COVID19_CaseReport.csv" comprises 4,000 rows (patients) and 102 columns (variables). These are hospitalized COVID-19 patients who were followed up until Aprtil 4, 2020. The set of all variable names in the data set are categorized below as follows:

- **Personal information**: "caseid" (Patient ID in the CHESS data), "trustcode" (NHS trust code), "trustname" (NHS trust name), "dateupdated", "weekno", "weekofadmission", "yearofadmission" (Timing information on hospital admission), "ageyear" (Patient age in years), "agemonth" (Patient age residual in months $< 12$), "ethnicity" (Patien ethnicity), "postcode", "sex", "estimateddate" (Estimated date of onset of symptoms), "notknownonset" (Indicates if onset of symptoms is unknown), "obesityclinical", "obesitybmi" (Obesity information), "pregnancy", "gestationweek", "prematurity" (Pregnancy information),"travel", "travelto", "traveldateofreturn", "travelin14days", "travelin14dayscondition" (Travel information), "worksashealthcareworker", "contactwithconfirmedcovid19case", "contactwithconfirmedcovid19casec".

- **Laboratory details**: "infectionswabdate", "labtestdate" (Dates at which swab was taken and lab test conducted), "typeofspecimen" (Type of swab, e.g. nasal, etc), "otherspecimentype", "covid19" (Whether the test result was positive for COVID-19), "influenzaah1n1pdm2009", "influenzaah3n2", "influenzab", "influenzaanonsubtyped", "influenzaaunsubtypable", "rsv", "otherresult" (Whether test result was positive for H1N1, H3N2, RSV or other influenza viruses).

- **Hospitalization details**: "hospitaladmissiondate", "hospitaladmissionhours", "hospitaladmissionminutes", "admittedhospital" (Timing for patient hospitalization), "admittedfrom" (Was the patient admitted from home or another location), "admissionflu", "admissioncovid19", "admissionrsv" (Reason for hospital admission), "ispneumoniacomplication", "is-

ardscomplication", "isunknowncomplication", "isothercoinfectionscomplication","isothercomplication", "issecondarybacterialpneumoniacom", "othercomplication" (Complications experienced in hospital), "dateadmittedicu", "hoursadmittedicu", "minutesadmittedicu", "dateleavingicu" (Date at which patient is admitted to or discharged from ICU), "sbother", "sbdate", "ventilatedwhilstadmitted","ventilatedwhilstadmitteddays", "patientecmo", "wasthepatientadmittedtoicu", "organismname", "daysecmo", "respiratorysupportnone", "oxygenviacannulaeormask", "highflownasaloxygen", "noninvasiveventilation", "invasivemechanicalventilation", "respiratorysupportecmo" (Respiratory support given to patient in ICU).

- **Antiviral treatment**: "anticovid19treatment" (Indication of whether a patient was given an antiviral drug).
- **Risk factors**: "chonicrepiratory", "asthmarequiring", "chronicheart", "chronicrenal", "chronicliver", "chronicneurological", "isdiabetes", "immunosuppressiontreatment", "immunosuppressiondisease", "hypertension", "othercondition" (List of all patient comorbidities).
- **Outcome**: "finaloutcome" (Patient status), "finaloutcomedate" (Date at which outcome is recorded), "transferdestination", "causeofdeath".

### 1.1.2 Variables involved in model development

Among the variables above, we used personal information and comorbidities to develop version 0.1 of the model. Personal information were limited to age, gender, obesity and pregnancy. Comorbidities included the variables: "Chronic Respiratory", "Asthma", "Chronic Heart", "Chronic Renal", "Chronic Liver", "Chronic Neurological", "Diabetes", "Immunosuppression Treatment", "Immunosuppression Disease", "Other Comorbidities", "Hypertension".

## 1.2 Data processing steps and assumptions

The initial data set contained records for 4,000 patients. We applied the following inclusion criteria for patient data used to develop our model. First, we excluded 15 patients who had missing hospitalization dates since for these patients we do not know the follow up times associated with their final outcomes. Next, we excluded 134 patients who had no confirmed COVID-19 diagnosis in the column "covid19". The flowchart for patient inclusion is given in Fig. 1.

We adopted the following assumptions when clearning the data set. First, a missing field in the comorbidities columns means that the patient had no diagnosis for such comorbidity. We also assumed that patients with age 0 years

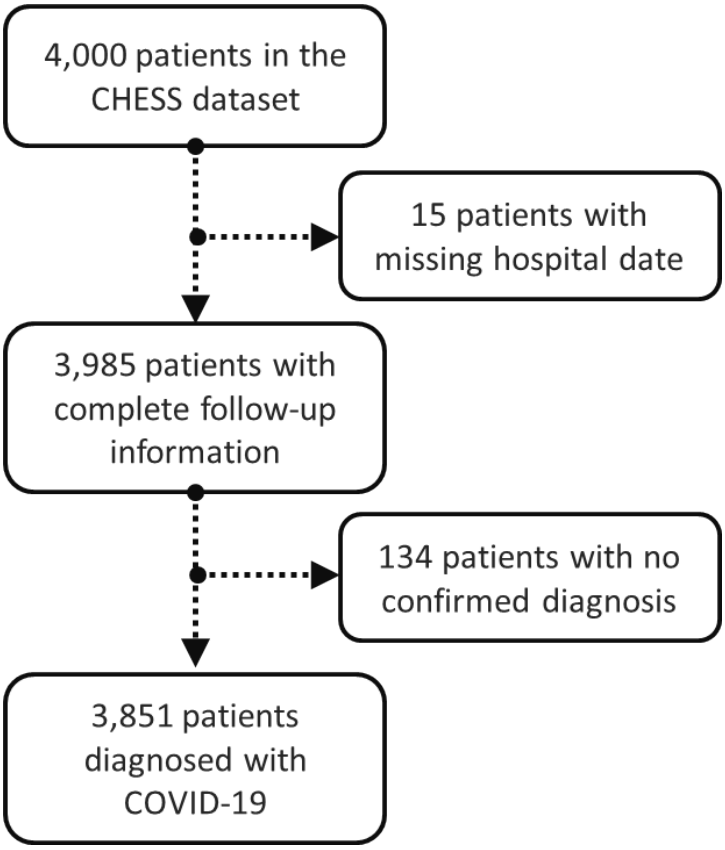


Figure 1: Flow chart for patient inclusion criteria.

and 0 errors have missing values for their age variable. We used the multiple imputation by chained equation (MICE) model in Python's `sklearn` library to impute the missing variables.

Functions for data processing are available in the `data-processing.py` file in the `data` folder. In this file, the function `get-data(curr-date)` takes as an input the date for the latest data feed. This function invokes the function `clean-CHESS-data` which applies the inclusion criteria described above and processes the variables to be input to the machine learning models.

### 1.3 Data statistics

Among all hospitalized patients by April 4, 2020, the death and discharge rates where 12% and 13 % respectively, with the remaining 75% of patients still hospitalized. Patients admitted to the ICU were 40% of the total hospitalized patients. Outcomes and trajectories of patients are depicted in Fig. 2. The median age of hospitalized patients was 68.8 years and 37% were female.

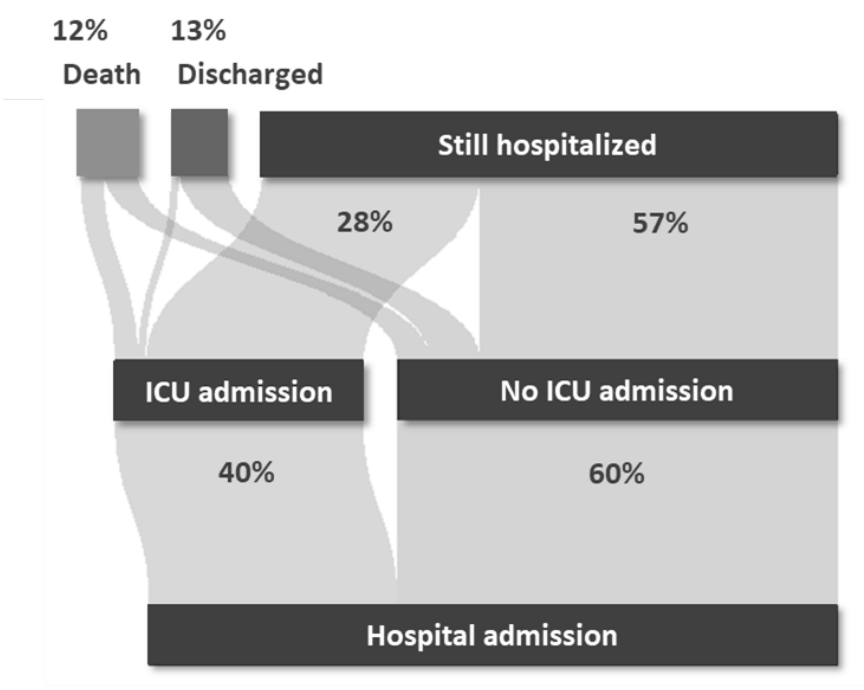


Figure 2: Fraction of patients in different outcome trajectories.

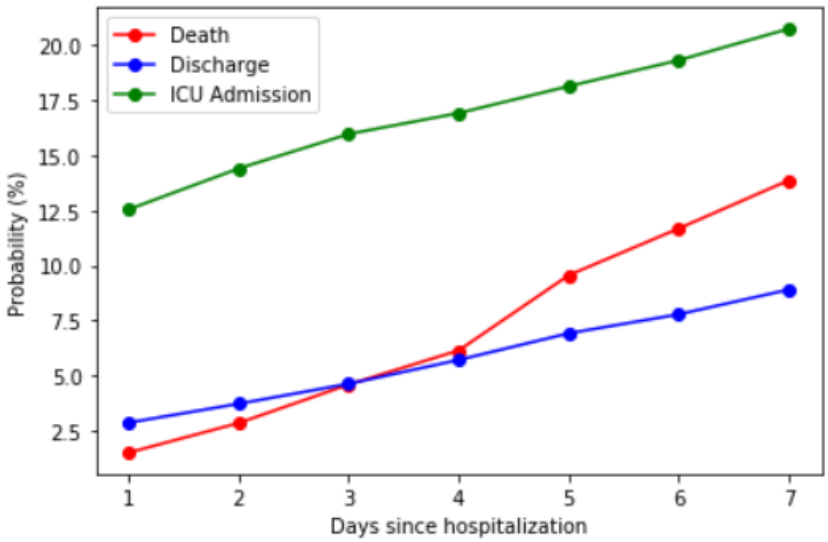


Figure 3: Exemplary model outputs for a single patient.

# 2 Model Development

In this Section, we describe the steps involved in developing the machine learning model for predicting individual-level probability of death, discharge and ICU admission at hospitalization time.

## 2.1 Modeling assumptions

We assume that the predictions are made at hospitalization time for each patient. At this time, the information available about the patient are their personal information ("Age", "Gender", "Obesity" and "Pregnancy") and comorbidities ("chonicrepiratory", "asthmarequiring", "chronicheart", "chronicrenal", "chronicliver", "chronicneurological", "isdiabetes", "immunosuppressiontreatment", "immunosuppressiondisease", "hypertension", "othercondition"). All input variables to the model are binary except for the age variable.

## 2.2 Machine learning model development

The machine learning model predicts patients' death, discharge and ICU admission probabilities over a future time horizon of 14 days. Fig. 3 depicts exemplary model outputs for 1 patient.

The model is developed through the following steps (depicted in Fig. 4). First, for each of the three predicted events (death, discharge and ICU admission), we create corresponding data sets with binary outcomes for each patient at each of the 14 prediction time horizons $\{1 \text{ day}, \ldots, 14 \text{ days}\}$. For each time horizon $T$, these data sets are created by including all patients who have follow up for $T$ days, and then binary outcomes are creates as follows: if the patient experienced the outcome by time $T$, a label 1 is attached to the patient, otherwise a label 0 is attached. The process is then repeated for all time horizons in $\{1 \text{ day}, \ldots, 14 \text{ days}\}$. After obtaining the probability curve over all the 14 time horizons, the final probability curve is smoothed via a piece-wise monotonic interpolation.

To capture uncertainty, we repeat the process above $N$ times (using bootstrapped subsets of the data set) to obtain $N$ versions of the model. The final prediction is obtained by averaging the predictions made by the $N$ models.

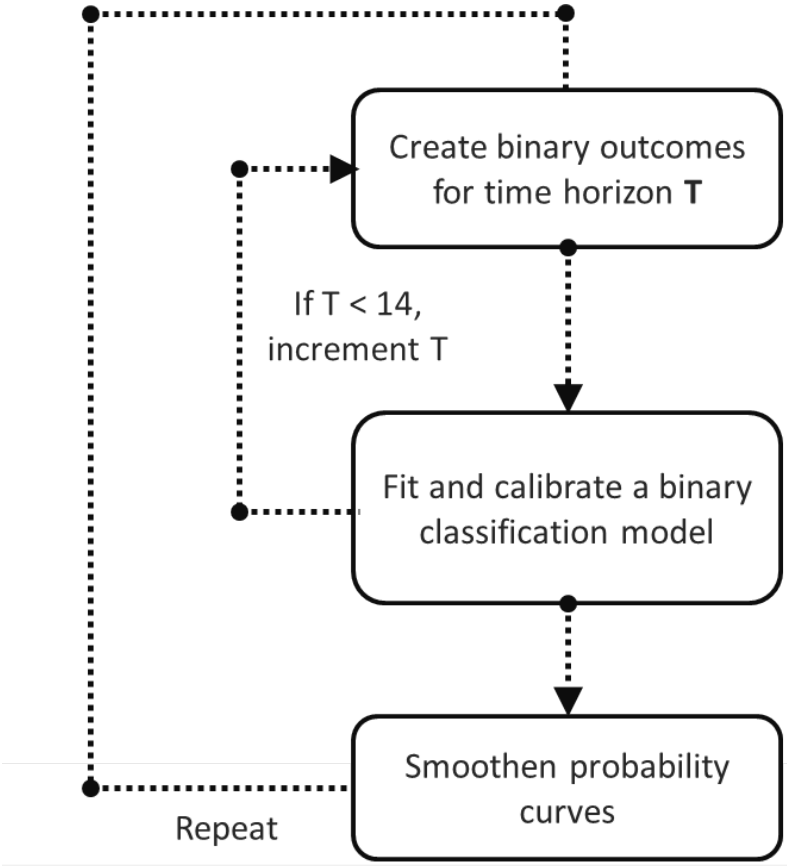


Figure 4: Model development steps.

## 2.3 Applying the model to the CHESS data

The current version of the model is applied to the CHESS data including only age, gender, obesity, pregnancy and comorbidity information for patient-level features. Information on ethnic background, social deprivation or geographic locations are not currently considered in the model.

## 2.4 Use of augmented data sets

Model updates will include augmented data linkages (e.g., hospital episode statistics) through which information on prescriptions, medical history, and ethnicity can be derived. To prevent any modeling with respect to protected or sensitive characteristics (such as race/ethnicity), the following precautions are taken:

- Protected characteristics will be incorporated in the model only if they prove to be predictive of patients' prognoses. This will be tested through a log-rank statistical test of differences in survival curves across patients stratified by different values of the protected characteristic. For instance, ethnicity information will be used as predictors in the model only if different ethnic groups display significantly different outcomes.
- Model validation will be conducted on a group-level rather than a population-level. If ethnicity information (or any other sensitive information) are added, the model will be trained to optimize accuracy and calibration metrics across each ethnic group.
- Potential hidden biases in assignments of interventions or ICU admissions will be accounted for using propensity-weighting techniques that ensure that interventions are not confounded by protected characteristics.