READ ME – Process_Data_Table_Age

This document describes the purpose of the process data code and describe the function of each cell. Read this after reading **README – Input_data**.

load packages – load the required packages

Define dates – this defines the dates required for the code. "file_date" is a date indexing for the input and output files. "dtpull" is the final date of admissions included in the input data. When extracting the input data, we recommend removing any admissions occurring on the last two days. This is because of delays from swabbing until results. Therefore, dtpull should generally be two days prior to the current date. "dtcensor" is the actual date on which the data is extracted. This variable tells the model how long people who are still in hospital have been in the hospital for.

Load data – this loads the latest data extract. This should be of the form:

["NHSNumber", "HospitalAdmissionTime", "StartTimeCriticalCare", "DischargeTimeCriticalCare", "HospitalDischargeTime", "DateOfDeath", "DateOfBirth"]

Make sure to remove headers from the data, as the code will add the required column headers. Here we require dates in the format "yyyy-mm-dd hh:mm:ss". Hours and minutes are important (if possible) since some events might occur on the same day. Synthetic input data is provided in *Simulated_COVID.csv*. You need to be careful with this date format, as sometimes when opening the csv in excel, the date format automatically changes. In this case, the dates will need to be changed back to this format using the custom cell formatting option in excel before saving the file.

 # Convert string dates to datetimes – convert dates to the required dates format

 # For each NHS Number, extract the patient's first admission and final discharge (can be omitted if interested in tracking multiple admissions) – this cell deals with duplicate entries and multiple ICU admissions. First, if a patient has a critical care admission time, we remove any duplicate rows for this patient without a critical care admission time. This if a patient is admitted and discharge from critical care multiple time, we only record the first entry and final discharge and take this a single critical care stay, removing all duplicate rows. We repeat this for hospital admission and discharge, so that each patient has a single admission time and single discharge time. If any discharge time is *nan* then the patient is assumed to still be in hospital and *nan* is taken as the final discharge time is *nan.* Finally, if any patient dies within five days of discharge, we change the discharge date to be date of death, since it is highly likely this is a COVID-19 death and that the patient was transferred elsewhere (for example a care home) to die. If the user would prefer not to include these deaths, the line :

(DeathDate < last_discharge+timedelta(5))

can be changed to:

(DeathDate < last_discharge+timedelta(0))

Matrix of allowed transitions – define the allowed transitions. State 1 – acute ward, State 2 – ICU, State 3 – stepdown ward, State 4 – discharge, State 5 – death

Create a data frame – this creates a data frame containing the length of stay in each state for each individual. This includes the length of time before an individual moves to their next state. This also adds censored rows for the transitions that did not occur, since in this case we only know that the transition time was longer than the transition time of the observed event. This data frame

indicates the state number of each individual at a time relative to their hospital admission time, and the potential transition time that they can make from that state. For each individual, an age indicator is added, taking four age groups: 0-25. 25-50, 51-75, 76+. This is to allow age to be used a covariate in predicting length of stay/outcome. A message is returned *"Negative or zero length duration encountered, please check rounding in data"* if a negative duration is encountered. If this is the case, check that all event time have hours and minutes.

Save length of stay data frame – save the above data frame to a csv file

Define time series for census – This defines the time series that we are interested in calculating the hospital census over

Create a 'census' of the hospital – Based on the processed data frame (i.e. with duplicates removed and discharges/critical care stays corrected), this cell calculates the number of individuals in each state for each date in the time series specified above. A census is created for each of the different age groups. This allows the dynamics of different age groups over time to be studied independently.

Save census – This saves the census data to a csv file.

Generate input file for Fit_Admissions_Growth.R – This generates the input file for *Fit_Admissions_Growth.R*. This records the number of daily admissions each day for the last *growth_len* days for each age group (and the total daily admissions). We recommend using a *growth_len* between 28 and 100.