Appendix 5. Documentation for data clean up for hospitalizations

The following steps were used to ensure those who received ventilator use or ICU care are counted for in the hospitalizations, and to calculate the length of stay for individuals requiring hospitalization.

The variables in **bold** were created, whereas other CAPITALIZED variables are from the original data source.

1. Create new variable **HOSPSTART**: if HOSP = 1, take earliest date from: HOSPADMITDATE, HOSPDISCHARGEDATE, ICUSTART2, else blank. If there are no dates then show a blank.

   The biggest problem with this data field was the HOSPADMITDATEs were frequently in the future of the HOSPDISCHARGEDATE so we needed to make sure chronologically they made sense (the majority).

2. Create new variable **HOSPEND**: if HOSP = 1, take latest date from: ICUEND2, HOSPADMITDATE, HOSPDISCHARGEDATE, else blank. If there are no dates then show a blank.

3. Create new variable **HOSPCHECK**: if HOSP = 1, HOSPSTART – HOSPEND = 0, and count(HOSPADMITDATE + HOSPDISCHARGEDATE) = 1, then 1, else 0. Same rationale as the other VENTCHECK and ICUCHECK.

4. No need to create **HOSPEND2** because there is nothing left to adjust to at this stage. For those with no discharge date, there is no assumption that can be made here.

   For all calculations, we only used infected individuals who are considered 'Resolved' or 'Fatal' in the OUTCOME1 field so the analysis would not be affected by censoring.

5. Create new variable **HOSPSTART2** for those who are missing HOSPSTART (and subsequently HOSPADMITDATE). If HOSPCHECK =1, COUNT(HOSPADMITDATE) = 0, COUNT(SYMPTOMONSETDATE)=1, AND SYMPTOMONSETDATE < HOSPDISCHARGEDATE, then = SYMPTOMONSETDATE, else HOSPSTART. This assumes that for those with a hospitalization, and a discharge date, that their admit date is time of symptom onset if they had no admission date *(assumption).*

6. Create new variable **WARD_LOS**. If HOSPEND-HOSPSTART2 < = 0 and HOSPCHECK = 1, then "X", else if HOSPSTART2 < ACCURATE_EPISODE_DATE, "X", then HOSPEND-HOSPSTART2. The else if statement of code is to correct for HOSPADMIT that were erroneously coded, or before the ACCURATE_EPISODE_DATE.

   If the person had no date fields completed but were considered hospitalized, they were counted in the proportions but excluded for LOS analysis.

7. Created two variables **DELAY_Symptom:** if HOSP = 1, Count(SYMPTOMONSETDATE)=1, and HOSPSTART2 >= SYMPTOM_ONSET_DATE, then HOSPSTART2 – SYMPTOM_ONSET_DATE, else "NA". **DELAY_Accurate:** The same was done but using ACCURATE_EPISODE_DATE instead of SYMPTOM_ONSET_DATE.

   The else if statement was to remove those with dates that were incorrectly entered into the HOSPADMITDATE field of 2018, 2019, early 2020 (could be errors) .