Appendix 3. Documentation for data clean up

The following steps were followed to ensure those who used acute care resources are captured. For example, someone who required a ventilator/intubation is counted as 1, and, by definition, required ICU admission. Those who needed ICU admission were, by definition, hospitalized. We assumed that those who are intubated, are counted as individuals who required ventilator use, and to calculate the length of stay for individuals requiring ventilation.

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

1. Create new variable **VENT**=1 if INTUBATION or VENTILATOR = 'Yes'. This is needed to count those who required ventilation.

2. Create new variable **ICU2** = 1, if ICU = Yes or VENT= 1. This is needed for the ICU proportions data. Some who are ventilated or intubated, have nothing in the ICU data field ('Yes' or 'No') or their start/end dates.

3. Create new variable **HOSP** = 1, if ICU2=1 or HOSPITALIZED= "Yes". This is needed for the hospitalization proportions. Some who are on ventilator, or admitted to ICU have nothing in the HOSPITALIZATION data field or their start/end dates.

4. Create new variable **VENTSTART**: take earliest date from VENTSTARTDATE, INTUBSTARTDATE, VENTENDDATE, INTUBENDDATE. If there are no dates then show a blank.

5. Create new variable **VENTEND**: take latest date from VENTSTARTDATE, INTUBSTARTDATE, VENTENDDATE, INTUBENDDATE. If there are no dates then show a blank.

6. Create new variable **VENTCHECK**: This variable will show 1 if VENT = 1, VENTSTART – VENTEND = 0, and (count(*DATE) =1 or INTUBSTARTDATE=VENTSTARTDATE, else 0. The purpose of this variable is to flag if there is only one date in the four *DATE fields for intubation and ventilation, or if there are two dates but they are equal (e.g., INTUBSTARTDATE = VENTSTARTDATE) and this date is being used for the VENTSTART and VENTEND.

7. Create new variable **VENTEND2**: if ICU2=1, VENTCHECK=1, and count(ICUSTARTDATE, ICUENDDATE) > 1, take ICUENDDATE, otherwise VENTEND. The purpose of this variable is to check if there is a VENTCHECK flag, then see if there is an ICU end date from original dataset. If yes, then assume that for VENTEND2, if not then take the current VENTEND.

8. Create new variable **VENTSTART2**: if ICU2=1, VENTCHECK=1, and count(ICUSTARTDATE, ICUENDDATE) > 1, take ICUSTARTDATE, otherwise VENTSTART. The purpose of this variable is to check if there is a VENTCHECK flag, then see if there is an ICU start date from original dataset. If yes, then assume that for VENTSTART2, if not then take the current VENTSTART.

9. Create a new variable **VENT_LOS**: If VENTEND2-VENTSTART2 < = 0 AND VENTCHECK = 1, then "X", else VENTEND2-VENTSTART2. This variable is used for the LOS for patients on the ventilator. In the first part of that logic statement, it says if there is a flag, and the dates are not in the correct order, then LOS is not calculated and these individuals are excluded.