**Data structure for IPW analysis of attributable mortality:**

The required data structure will be a hospitalization-day data set (one line per person per calendar day).

Restrict to hospitalizations where:

1. hospital length of stay >=3 days
2. there is a measurement on first or second day of admission for all time-varying confounders.
   1. If any of the following labs are not available from either hospital day 1 or day 2 then exclude the patient: WBC, HCT, Platelets, Sodium, Glucose, Creatinine. If missing day 1 impute with day 2 and vice versa.

For eligible hospitalizations, the data must have columns including:

1. Hospitalization id – indicates a unique hospitalization The same id number will be repeated on consecutive lines indexed by different Day indices for that hospitalization.
2. Day – corresponding to the follow-up day for a given id. The index for this can start at day 3 of hospitalization and it increases by 1 for each additional follow-up day. The max value of Day for a given id will be 63 (assuming we end follow-up at a max of 60 days post day 3).
3. A column for each time-fixed confounder (see list below). These values will be repeated for the same id on all rows with that id.
4. A column indicating “exposure” on a particular hospitalization-day: this will change for a given id depending on the value of the Day column and is an indicator of whether the candidate definition has been first met BY that day. This indicator should correspond to the first day of worsening oxygenation when a patient meets the NV-HAP definition. Please add indicators for each of def1, 2, 7, and 9. If a patient meets NV-HAP criteria more than once during a hospitalization only include their first episode for the purposes of this analysis. Carry forward this indicator such that once it switches to 1 on a given day-row for a given id, it stays 1 on all subsequent day rows for that id.
5. A separate column for each time-varying confounder that may change each day. For each row for day t, include a column with each confounder value for day t-2 . Values on a given line will be the last measured value relative to day t-2. This will be carried forward from last line if no new measure on day day t-2, otherwise it will be the new measure on that day. We can also include indicators of time since la st measurement relative to day t-2 for a line indexed by day t. See list and notes below. Include a column to indicate if the patient is on a ventilator by day t using both oxygen device type and mechanical ventilation ICD10s; code this variable as 0 until it switches to 1. If a patient comes off a ventilator then revert to 0. If there is no oxygen device indicated or mechanical ventilation code for a given day then carry forward the last known value.
6. A column for hospital mortality status on the current day. On the first day that this variable switches to 1 this is the last record in the data set for that patient.
7. A column for discharge status on the current day. On first day this switches to 1 this is the last record for that encounter in the data set. Mortality and discharge are competing events so only one of these indicators can be 1.
8. I am assuming there is no loss to follow-up here in the sense that we know fully status of discharge timing and mortality within hospital up to 60 days from day 3 of admission. If there is, let me know and will add instruction on this.
9. SiteID: A column indicating which hospital within HCA (or VA).

Output of analysis:

1. Cumulative incidence (risk) curves over 30 and 60 days follow-up,beginning from day of admission\* under the different hypothetical interventions considered separately for outcomes mortality and discharge. (so point estimates for all follow up times)
2. Risk differences/risk ratio point estimates comparing interventions for both outcomes plus 95% Cis for day 30 and 60, beginning from day of admission (CIs obtained via bootstrap, per individual patient rather than per hospitalization)

* Because the data is restricted to those with at least 3 hospital days, the counterfactual contrast we are going after is conditional on surviving and remaining hospitalized to day 3, and therefore (by definition) the risk of either of these events is 0 on days 1 and 2 of hospitalization. Even in a trial to interpret this conditional contrast as a causal effect we have to assume that the treatment/exposure does not affect either discharge or survival in this short pre-day 3 time period (which maybe is not unreasonable).

Combining VA and HCA results:

We will take a weighted average of the estimates from VA and HCA based on the number of eligible hospitalizations in each. Specifically if n1 is number in VA and n2 in HCA then the point estimates will be

Estimate(VA)\*(n1/n1+n2)+ Estimate(HCA)\*(n2/n1+n2)

95% confidence intervals:

CIs obtained from a nonparametric bootstrap resampling hospitalization ids with replacement 200 times. Resampling will be based on the total n1+n2 ids, the algorithm will be computed in each sample at each site using only the ids selected from that site. Setting the same seed at both will ensure that the same split happens at both sites such that in both sets of code in bootstrap sample B there are n1(B) hospitalizations from VA and n2(B) hospitalizatiions from HCA – noting that n1(B) or n2(B) may include the same hospitalization from the original data selected more than once – such that n1(B)+n2(B)=n1+n2. Then the results will be combined in each bootstrap sample the same way they are combined for the estimates but replacing n1 with n1(B) and n2 with n2(B).

Interpretation:

We would like to interpret our estimates in terms of the effect of fully preventing NV-HAP in all hospitalizations in these systems compared to no intervention.

**Covariates:**

**Time-fixed confounders**

|  |  |  |
| --- | --- | --- |
| **Category** | **Covariate** | **Variable Name** |
| Hospital Features | number of beds (1-99, 100-199, 200-299, 300+), |  |
|  | teaching status (non-teaching vs teaching) |  |
|  | region (midwest, northeast, south, west). |  |
|  |  |  |
| Demographics | Age (included in propensity score model with 4 data adaptively selected knots) | age |
|  | Race (black, white, Asian, other, missing) | race |
|  | Sex | sex |
|  |  |  |
| Comorbidities | Diabetes (combine diabetes w and w/out complications) | diabetes |
|  | Cancer (combine lymphoma, solid tumor, and metastatic cancer) | cancer |
|  | Chronic pulmonary disease | cpd |
|  | Chronic kidney disease (renal failure) | renlfail |
|  | Chronic liver disease | liver |
|  | Congestive heart failure | chf |
|  | Cardiovascular disease (cardiac arrythmia, valvular disease, pulmonary circulation disorder, peripheral vascular disorder) | cardiovasc |
|  | Neurological disease (paralysis) | neuro |
|  | Obesity | obese |
|  | Alcohol abuse | alcohol |
|  | Weight loss | wghtloss |
|  | Deficiency anemias | anemdef |
|  | HIV and AIDS | aids |
|  | Rheumatologic disease (Charlson) | arth |
|  | Any hospitalization in past 90 days (indicator) |  |
|  |  |  |
|  |  |  |

**Time-varying confounders**

Severity of Illness

* Hospital service: cardiology, medical, surgical (plain and cardiovascular), neuroscience, oncology, , other. . Use lag 2 value in the dataset. Needs to be complete on day 1 or 2, can use last measured update. If service not available day 1 or 2, impute first available service to day 1 and 2, carryforward until service changes.
* ICU status on day t-2
* Laboratory tests (from the day *prior* to NV-HAP onset) – relative to each day of follow-up (starting from day 3 of hospitalization) we will want to include on each day line the relevant “history” that we think affects mortality and discharge beyond that day and also status on the candidate definition on that current day but avoiding any thing that is part of determining the definition given other variables in the exposure model, some more notes on this in example below. Include on each day line t (with first row containing measurements from hospital day 3 for all patients) last measured values of the following variables from two days prior (day t-2 relative to that line). Because these will not be measured every day, we can carry forward. If any of the following labs are not available from either day 1 or day 2 then exclude the patient: WBC, HCT, Platelets, Sodium, Glucose, Creatinine. If a given lab value is only available on day 1 or day 2 then impute the value from the available day (be it day 1 or day 2) to the other day.
* For ALT, Tbili, or Albumin provide for all days of follow-up , , with values from t-2 (if missing, leave as missing). Categorize as follows:

|  |  |  |
| --- | --- | --- |
| Lab | Result Range | Code as |
| ALT | Missing | 0 |
|  | 0-50 | 1 |
|  | 51-99 | 2 |
|  | 100-199 | 3 |
|  | 200-499 | 4 |
|  | ≥500 | 5 |
| Tbili | Missing | 0 |
|  | 0-1.0 | 1 |
|  | 1.1-1.9 | 2 |
|  | 2.0-2.9 | 3 |
|  | 3.0-4.9 | 4 |
|  | 5.0-9.9 | 5 |
|  | ≥=10 | 6 |
| Albumin | Missing | 0 |
|  | 0-0.9 | 1 |
|  | 1.0-1.9 | 2 |
|  | 2.-2.9 | 3 |
|  | 3.0-3.5 | 4 |
|  | ≥3.6 | 5 |

|  |  |
| --- | --- |
| Lab name | Variable Names |
| WBC | Last\_wbc, last\_wbc\_lag1, last\_wbc\_lag2 |
| HCT | last\_hct, last\_hct\_lag1, last\_hct\_lag2 |
| Platelets | last\_platelets, last\_platelets\_lag1, last\_platelets\_lag2 |
| Sodium | last\_na, last\_na\_lag1, last\_na\_lag2 |
| ALT | last\_alt, last\_alt\_lag1, last\_alt\_lag2 |
| Tbili | last\_bilirubin, last\_bilirubin\_lag1, last\_bilirubin\_lag2 |
| Glucose | last\_glucose, last\_glucose\_lag1, last\_glucose\_lag2 |
| Creatinine | last\_creatinine, last\_creatinine\_lag1, last\_creatine\_lag2 |
| Albumin | last\_albumin, last\_albumin\_lag1, last\_albumin\_lag2 |

* On the line indexed by day t, please also create a variable for time since any last lab measure for **core labs only** (WBC, HCT, Platelets, Sodium, Glucose, Creatine), on day t-2: This takes the value zero if there was a current lab measure on day t-2. Otherwise it is the number of days since any new lab was taken relative to day t-2. So, for example, say for a given hospitalization a person only received a new lab on days 2 and 5. Then on the line indexed by day 4 this variable would take the value 0, on day 5 it would be 1, day 6 it would be 2 and on day 7 it would reset to 0 (since there was a new measure 2 days ago on day 5), on day 8 it would be 1 and so on.
  + Included as categorical variable 01-; 2-4; >=5 days
  + How to handle missing on day 1 or 2. Do imputation for day 1 and 2 (as described on page 1), and assume those were actual measurements for this variable.
* Baseline oxygenation (from the day *prior* to NV-HAP onset) – this should be counted like the other variables above (last measurement of oxygenation on day t-2). If no oxygenation data available for a day carry forward last available value.
  + Variable names for O2 sat for day, lag of 2 days: last\_spo2, , last\_spo2\_lag2
  + Categories: >= 95, 90-94, and less than 90

Add oxygen device variable names for day, , lag of 2 days

* baseline\_oxygen, , baseline\_oxygen\_lag2
  + Use an ordinal scale. Use first device of the day for this variable.
    - None with O2 sat >=95% = 0
    - None with O2 sat <95% = 1
    - Nasal cannula = 2
    - Simple mask = 3
    - Oxygen conserving device = 4
    - Non-rebreather = 5
    - High-flow = 6
    - BIPAP = 7
    - Ventilator = 8 (propensity score models will be restricted to records with lag 2 of this variable not equal to 8)

If need to combine oxygen device categories can combine 0 and 1; 2; 3, 4 and 5; 6 and 7

An example of the data structure for one id who first meets the candidate definition for exposure on day 6 of hospitalization and dies on day 9 prior to discharge. The “ts” variable is time since last measurement of WBC on day t-1. In this example, ts=0 on day =3 means they had a WBC measure on day 2 which was 3500, this is carried forward until line day =6 which indicates that a new measurement of 3900 was available on day 5.

Variable names:

personid(combination of hospital id and admittance id), id(sequence # of patient with

multiple admittance ids),

day(t) = dayt,

exposed(t) = exposed, exposed\_lag1, exposed\_lag2

ts WBC(t-1) = ts

ventilator = ventilator, ventilator\_lag1, ventilator\_lag2

discharge = discharge, discharge\_lag1, discharge\_lag2

death = death, death\_lag1, death\_lag2

ventilator = ventilator, ventilator\_lag1, ventilator\_lag2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Obs | ID | day(t) | exposed(t) | ts | WBC(t-1) | race | ventilator | discharge | death |
| 1 | 66 | 3 | 0 | 0 | 3500 | white | 0 | 0 | 0 |
| 2 | 66 | 4 | 0 | 1 | 3500 | white | 0 | 0 | 0 |
| 3 | 66 | 5 | 0 | 2 | 3500 | white | 0 | 0 | 0 |
| 4 | 66 | 6 | 1 | 0 | 3900 | white | 0 | 0 | 0 |
| 5 | 66 | 7 | 1 | 1 | 3900 | white | 0 | 0 | 0 |
| 6 | 66 | 8 | 1 | 2 | 3900 | white | 0 | 0 | 1 |

+++++