EDA Question on MIMIC-III data: What factors might lead to 30-day readmission?

Kaijie Zhang (kaz029@ucsd.edu)

30-day readmission is an important indicator for assessing the quality of medical services, and many insurance companies also pay attention to it. In an LLM evaluation paper we previously read, there was an attempt to compare an LLM's accuracy on standard diagnosis with its accuracy in identifying reasons for 30-day readmission; the latter was often much lower than the former. The causes of 30-day readmission may not depend solely on information in the diagnostic report and may be influenced by multiple factors. Therefore, I would like to explore the related factors.

Result & Conclusion

The risk factors for 30 day readmission that we discuss are largely intuitive, yet they are diverse and complex.

Age:

Older patients generally face a higher risk of readmission. Advanced age often comes with multimorbidity and frailty, making post discharge complications more likely.

Length of stay:

A longer hospitalization, although the effect rises only within a range, is associated with a higher risk of readmission. Prolonged stay signals greater illness severity and dependence on care, which accords with clinical common sense.

Disease category:

Specific diagnoses are strongly linked to high 30-day readmission. The top signals in the figure include anaerobic septicemia, acute alcoholic hepatitis, acute respiratory failure after trauma or surgery, acute diastolic heart failure, pneumonia NOS, and major cardiovascular or renal complications. These conditions share a high risk of relapse or unresolved issues, so they warrant closer follow up after discharge.

Discharge to facility:

Home level care is associated with a high burden of 30 day readmission. Discharge to rehabilitation hospitals, skilled nursing facilities, or other non home settings also carries a substantial risk. This underscores the importance of the destination after discharge, likely because it closely reflects care quality; shortfalls tend to lead to return admissions. Also note that some long-term hostilities have high rates in emergency cases, perhaps because the patient's illness is too severe to be discontinued from good medical care.

Diagnostic reports:

The effect of diagnostic reports on 30 day readmission was not examined in depth in this report. Because the notes were vectorized with TF-IDF, the identified features depend heavily on specific clinical terminology. This points to further directions such as modeling particular combinations of diseases, severity markers and key numerical values that are mentioned in notes and can also be extracted from chartevents, and special remarks that imply residual disease.

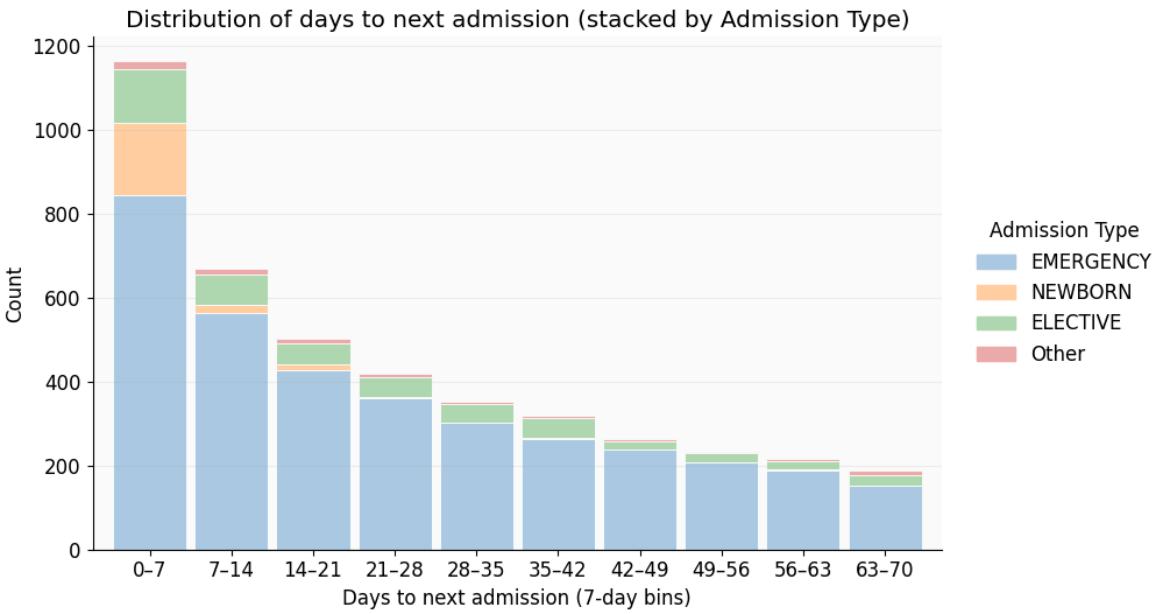
Poor social support:

As hinted by the discharge disposition findings, this commonsense factor merits separate study by linking to patients' socioeconomic circumstances.

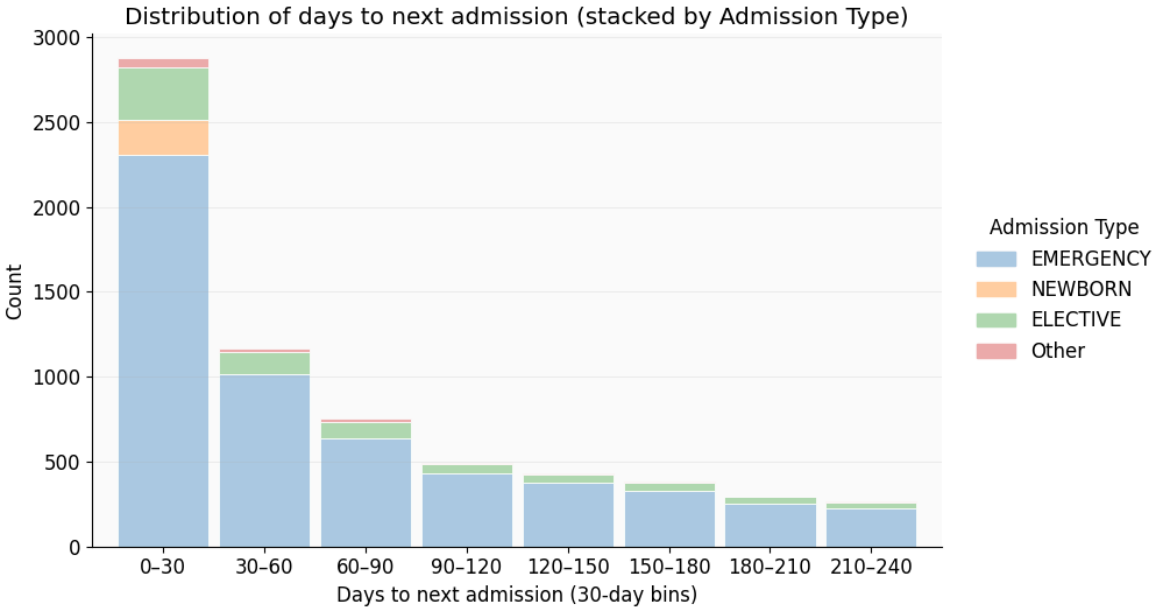
Else to say:

With the features reviewed so far, a usable predictive model is already feasible. Given the multifactorial nature of 30 day readmission, we should first study these features thoroughly, then attempt to build a model that approaches state of the art performance.

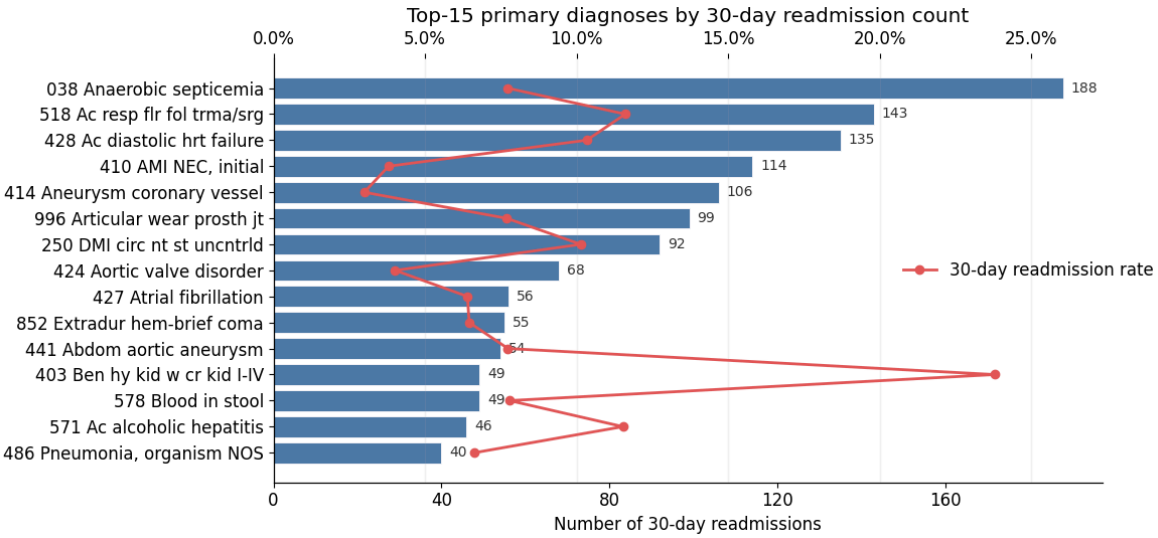
Basic EDA and Report



As the number of days increases, readmission cases decrease. Regardless of the interval, ICU entries via EMERGENCY account for the vast majority; therefore, EMERGENCY is a key indicator. We also note that NEWBORN cases typically return to the ICU only within the first 21 days, indicating a strong association between obstetric sequelae and readmission; this can serve as a reference.

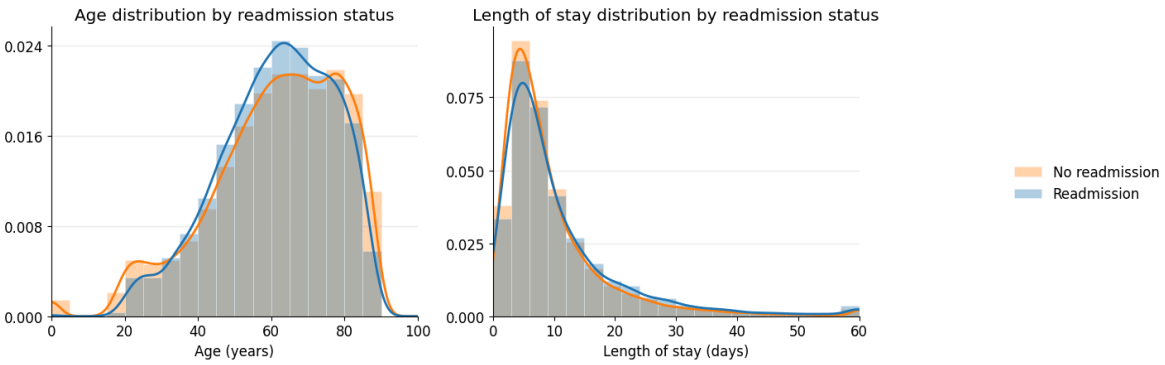


The plots show similar trends within 70 days. Admissions occurring after 240 days exhibit weak correlation with prior events, so they are not prioritized in this visualization.

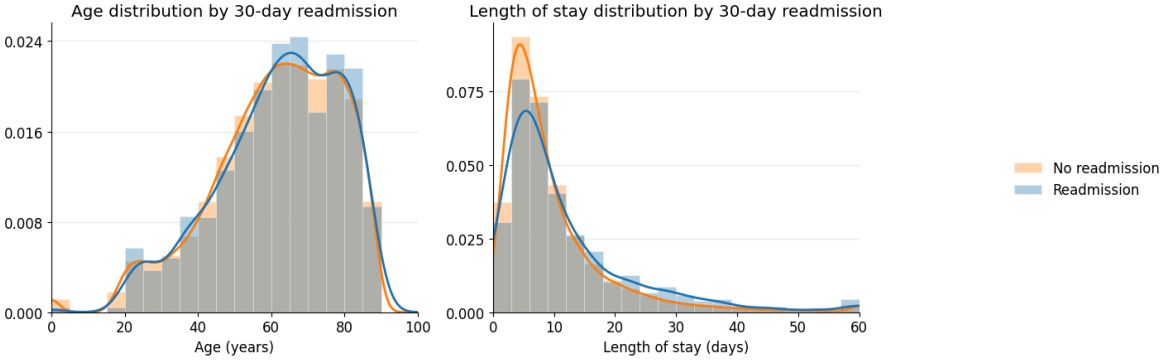


We observe that among the top K high contribution diagnoses, a large number of readmissions does not imply a high readmission rate. In the top 15, the two measures are sometimes even reversed. Overall, certain conditions stand out (such as acute alcoholic hepatitis, having rate > 20%), which shows an exceptionally high 30 day readmission rate. This suggests a strong association between specific diagnoses and 30 day readmission.

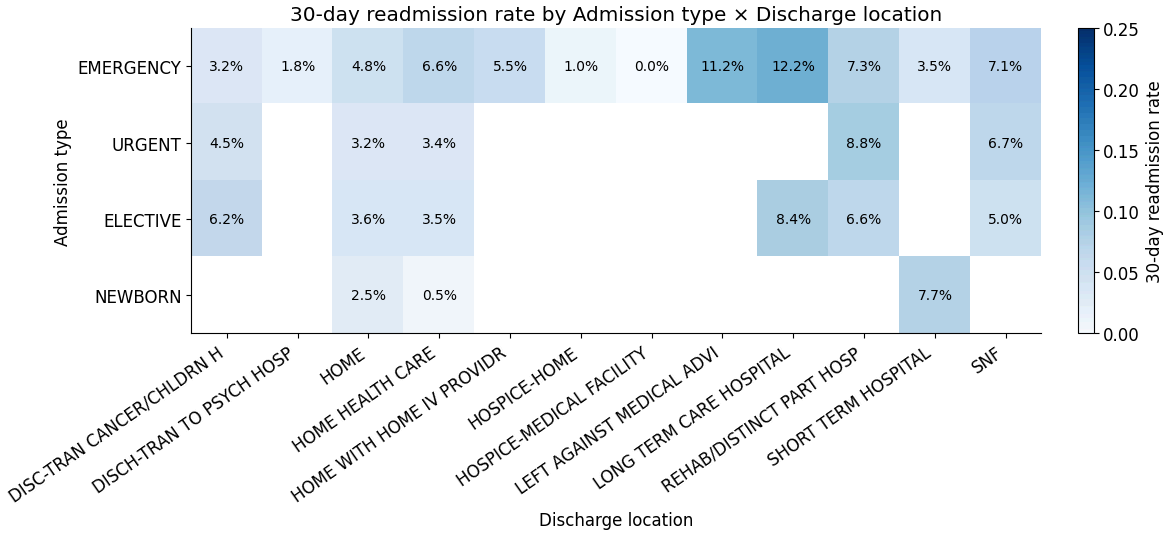
```
Rows: 53122 | AGE_DAYS valid: 51055 | LOS_DAYS valid: 53122 | any readmit: 10966
| 30d readmit: 2933
```



Having a prior readmission does not change the overall shape of the distributions, but the absolute rates are generally slightly higher in the prior readmission group. In the early ranges the group without prior readmission is higher, while in the later ranges the prior group overtakes it. Rates are consistently higher among patients aged sixty and above, and among those whose hospital stay exceeds three days but is under fifteen days. This suggests that several days in the ICU without discharge often signal greater severity and a higher probability of return after discharge. When age and length of stay reach very high values, the observed rates decline, which may reflect smaller sample sizes and mortality.



The 30 day readmission rate shows a distribution similar to the overall readmission rate, so the same conclusions apply. Notably, among patients without any prior readmission, the rate for those aged zero to ten years and for short stays exceeds that of patients with a prior readmission. This pattern is likely driven by newborn cases related to delivery, where there were no earlier readmissions related to the current episode and the length of stay tends to be short, although this requires confirmation.



We already knew that EMERGENCY accounts for the largest share of readmissions. Building on this, we observe a strong association between the 30-day readmission rate and discharge disposition. After patients are stepped down from the ICU and discharged to HOME or HOME HEALTH CARE, there are substantial numbers of 30-day

readmissions across these admission-type groups. This suggests that home-level care, potentially due to quality gaps, struggles to bridge the post-ICU transition. It also implies a possible explanation: some patients may have poor social support: poverty, living alone, and lack of follow-up caregivers may increase readmission risk (so within HOME and HOME HEALTH CARE we should focus on these subgroups for further exploration). Also note that some long-term hostilities have high rates in emergency cases, perhaps because the patient's illness is too severe to be discontinued from good medical care.

Moreover, REHAB/DISTINCT PART HOSP, SHORT TERM, and SNF (Skilled Nursing Facility) also show considerable 30-day readmissions, indicating that discharge to non-home settings such as rehabilitation hospitals or nursing facilities is likewise associated with nontrivial readmission risk.

```
Admission Date:  2119-5-4              Discharge Date:   2119-5-25

Service: CARDIOTHORACIC

Allergies:
Amlodipine

Attending:Last Name (NamePattern1) 1561
Chief Complaint:
81 yo F smoker w/ COPD, severe TBM, s/p tracheobronchoplasty 5-5
s/p perc trach 5-13

Major Surgical or Invasive Procedure:
bronchoscopy 3/31,4/2,3,6-12, 5-17, 5-19
s/p trachealplasty 5-5
percutaneous tracheostomy 5-13 after fa

... [content truncated] ...

ion-> head injury & rib fracture.
TBM- s/p tracheoplasty.

Discharge Condition:
good

Discharge Instructions:
please update Dr.Name (NI) 1816 Telephone/Fax (1) 170 office for:  fever,
shortness of breath, chest pain , productive cough or if you
have any questions or concerns.

Completed by:2119-5-25
```

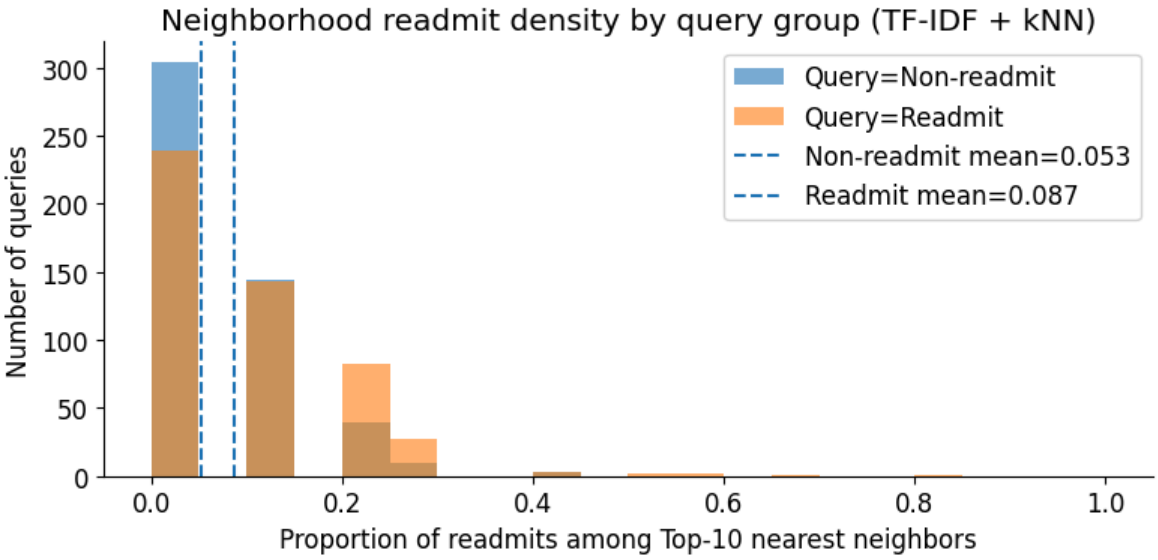
Example note/report for later use.

```
Available discharge summaries: 47463
  Readmit=1: 2778
  Readmit=0: 44685

Corpus shape: (47463, 50000) (docs × features)

For each query we take Top-10 neighbors; below is the average fraction of 30-day
readmits among neighbors:
Query=Readmit group     mean neighbor readmit fraction = 0.0874
Query=Non-readmit group mean neighbor readmit fraction = 0.0528
Difference (readmit – non-readmit)                     = 0.034600000000000006

Bootstrap 95% CI:  Readmit   0.087 [0.078, 0.097]  |  Non-readmit 0.053 [0.046,
0.060]
```




```
Permutation test: Δ_obs=0.0346,  p-value=0.000050
```

Note: Neighborhood - Most relevant/similar (TF-IDF) notes/reports

Note: Query - notes/reports; would be defined as query to retrive while tuning LLMs

Using 47,463 discharge summaries, I embedded the text with TF-IDF bigrams and used cosine kNN to find the Top 10 nearest neighbors for each admission while explicitly excluding the query note and all notes from the same patient. The average fraction of 30-day readmissions among neighbors was 0.087 for readmitted queries versus 0.053 for non-readmitted queries, a difference of about 3.4 percentage points, with bootstrap 95 percent confidence intervals of [0.078, 0.097] and [0.046, 0.060]. It gives a p-value of 0.00005, which is much lower than 0.05 threshold. This shows that clinical notes and reports contain useful signal (most likely to be clinical/professional vocabs/combinations, due to the essence of TF-IDF bigrams) for distinguishing higher near-term readmission risk.

Conclusion is moved to the front.