Team 40: Predicting Mortality in ICU Setting (Project Draft).

Vincent La[*], Avi Ananthakrishnan[†]

November 11, 2018

Abstract

FILLER

---

[*]vincent.la@gatech.edu

[†]avinash.ananthakrishnan@gatech.edu

1   Introduction

Accurate knowledge of patient's clinical state is critical in a clinical setting. Patients in an intensive care unit are particularly in a critical state. Thus, if we can have better accurate mortality prediction within an ICU, we can use this to better understand which patients should be prioritized for clinical care and how we can better allocate resources. Within an ICU, there are many measurements being made from from various devices, but we know that there are both problems in terms of false positives and false negatives. While false negatives might seem to be more problematic, being too sensitive and introducing a lot of false positives also has problems as well. Too many false positives results in information fatigue and providers lose trust in signal as it's hard to separate from noise. By improving mortality prediction using data from EMRs and other clinically-relevant data, we can provide better predictions to improve efficiency and quality of care.

2   Background and Literature Review

Siontis et. al (2011) provide an empirical review of methodologies around mortality models. In this paper, the authors use Medline to identify studies published in 2009 that assessed AUC of predicting all cause mortality. What they find is that most studies at the time predicted mortality with modest accuracy, with large variability across populations, which makes sense since for since very high risk populations it may be easier to predict mortality, but for lower risk general populations, the class imbalance is very high with relatively few patients actually dying. A full list of the studies that the authors review and the approaches they use are documented in the paper (linked above). However, we also copy a portion of that table below as an illustrative example:

3   Methodology

What we will do in this project is to try to reproduce and improve models motivated by Unfolding Physiological State: Mortality Model in Intensive Care Units (Ghassemi et. al). In particular, what we find interesting is the use of free-text hospital notes to better predict mortality. This is because in many cases, there are already many mortality models, built both on Business Rules, and machine learning to predict mortality. Perhaps the biggest methodology to improve on existing mortality models is to use unstructured data, to see if there is any additional signal from notes and other text that we don't already get from structured data using EHRs.

To be more specific, we will use the same dataset as indicated in the Project guidelines, MIMIC III. This data set has 45,000 patients, 60,000 Hospital Admissions, 200,000 Clinical notes, and 6000 mortality events. We will then divide into 80% training, 20% test set. Similar to what is presented in the paper, we will first extract clinical baseline features, including age, sex, and SAPS-II score, from the database for every patient. We will then perform topic modeling, perhaps LDA, to do natural language processing on the notes to extract meaningful features. In terms of evaluating our model, we will predict on multiple

iterations of the outcome variable, just as the paper does. For example, to evaluate 24 hour mortality predictions, in the training set, we will create a 24 hour prediction window and using the data that is transformed into features, we will measure AUC of mortality where outcome variable is mortality within 24 hours. We will then test on held-out test data set to report final test AUC.

4   Experimental Results

4.1   Data Exploration

The dataset contains 58,976 distinct hospital admissions with 46,520 distinct patients. Thus, most people in this dataset were admitted once, which makes sense, but there is also a long tail of readmissions.

Figure 1: Admissions Distribution
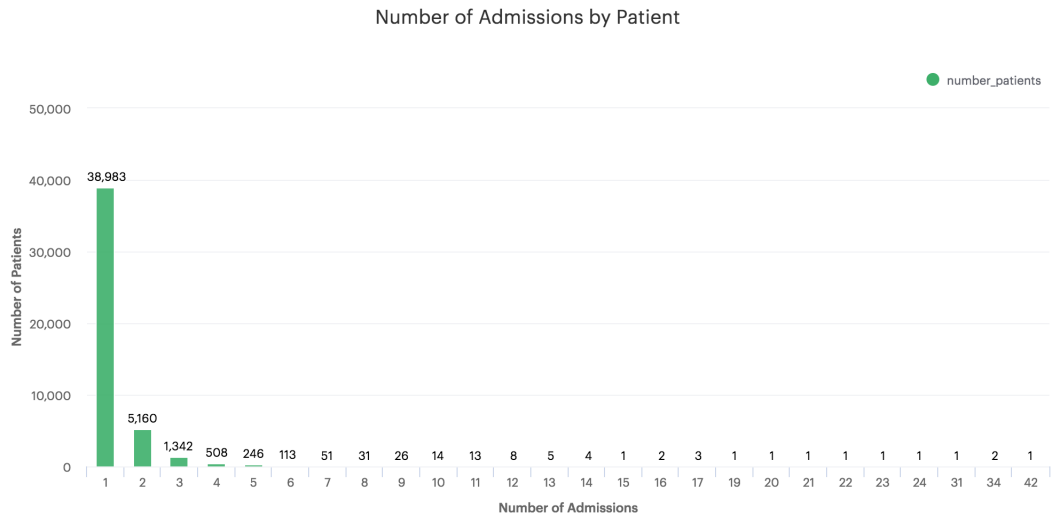
About 39,000 patients were admitted once, with 5,160 admitted twice. However, there s a long tail of about 2,000 remaining patients who have been admitted more than twice in this time period, with the maximum being 42 admissions over this time period.

Furthermore, we note that admission lengths on average are about 10 days, but the median length of stay is closer to 6.5 days. This makes sense as again, the distribution is likely right skewed. There are patients with very long outlier lengths of stays dragging mean up, but the median, which is less sensitive to outliers is lower.

Figure 2: Admissions Lengths (In Days)

	minimum	quartile_25	mean	quartile_50	quartile_75	maximum
1	-0.945	3.744	10.1084081321	6.467	11.795	290.66

4. Admission Lengths (In Days)

Note that there are a few patients with negative lengths of stays. This is likely bad data. The longest a patient was in the hospital is about 290 days which is almost 10 months.

Next, we look at the breakdown of admissions by patients' insurance. This is interesting because the type of insurance a patient has is likely to be very correlated with patient characteristics. For example, Medicare patients will likely be older and have more acute risk. Medicaid is likely correlated with meaningful socioeconomic characteristics. As expected, many of the patients are Medicare patients, at about 48% of the total population. This makes sense as Medicare patients are older and higher risk, so will comprise more of the admissions.

Figure 3: Admissions by Patients' Insurance

	insurance	number_patients	percentage_patients
1	Government	1783	0.03
2	Medicaid	5785	0.098
3	Medicare	28215	0.478
4	Private	22582	0.383
5	Self Pay	611	0.01

In addition, we note that about 5,800 patients are recorded as having died at some point after the admission.

Figure 4: Admissions Deaths

	death_after_admission	number_patients
1	false	42131
2	true	5813

In fact, if we look at the distribution of times to death, we see that the vast majority of deaths are recorded at the same time as the discharge time.

Figure 5: Admissions Deaths Time Period

5. Admissions Death Compared to Discharge Time		
	time_period_of_death	number_patients
1	death at time of discharge	5789
2	death recorded post discharge	38
3	death recorded prior to discharge	27

Given this, it very likely means that mortalities recorded before the discharge time are due to clerical error. It's also possible that mortalities recorded after the discharge time are also a result of clerical error; although it is possible that the study followed up with the patient post-discharge.

4.2 Feature Extraction

Now that we have done some data exploration, this next section will discuss feature extraction and what features we actually used in our model to predict mortality.

5 Discussion

6 Conclusion

Filler

References

[1] Ghassemi M, Naumann T, Doshi-Velez F, et al. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. KDD. 2014;2014:75-84.

[2] Siontis GCM, Tzoulaki I, Ioannidis JPA. Predicting Death: An Empirical Evaluation of Predictive Tools for Mortality. Arch Intern Med. 2011;171(19):17211726. doi:10.1001/archinternmed.2011.334