Mining Unstructured Clinical Notes for Mortality Prediction

Vineet Kumar (19BM6JP46)
Rohit Bajpai (19BM6JP56)

June 10, 2020

Abstract

Unstructured clinical data such as nursing notes are very less used to build a predictive model for post-discharge mortality despite containing rich information. Our work examines a simple bag of words model approach for 7/30/180/365 day mortality prediction. We also explored syntactic sentiment dimensions from these nursing notes as a predictor of mortality, and report preliminary survival analysis results too. Our simple BOW model using XGBoost achieved 0.92 AUC for 30-day mortality.

1 Introduction

Wider adoption of Electronic Health Records (EHRs) in the hospital setting has given rise to a plethora of clinical data. EHR records patient demographic details, past medical history, periodic clinical measurements like patient vitals, test reports, medical interventions, and detailed clinical/nursing notes. Structured clinical (tabular) data contain a rich but incomplete picture of the patient. Clinical notes are recorded by attending nurses in free form textual format. They contain rich information relevant to the patient's response to treatment and illness trajectory as well. To better understand high-risk patients, health systems must leverage text analytics,to derive insights from free form clinical texts. NLP helps in interpretation of textual data. It can aid in information extraction, conversion of unstructured to structured data, Document categorization etc.However, due to their high free form nature utilizing these unstructured clinical descriptions (UCDs) in building clinical decision support systems is not much explored. Predicting post-discharge mortality is one of the major research areas in health-informatics Metersky et al. (2012).

1.1 Related Works

Recent advances in NLP and data mining has made it possible to mine these unstructured nursing notes for mortality prediction. Hence, there has been multiple studies in model enrichment Staff (2013), patient phenotyping Gehrmann et al. (2017), readmission prediction Shin et al. (2019) etc.

Moreover, the latest advances in deep learning technologies (DL) have encouraged the use of deep nets in the clinical domain as well, ranging from computer-aided diagnosis to genome sequencing Miotto et al. (2018). Classical ML & statistical techniques require feature engineering, which is sometimes very time consuming and requires domain expertise. In such scenarios, the DL approach has become very

useful; however, due to their black-box nature, the use of these sophisticated approach is still not much appreciated in clinical decision support systems (DSSs).

In this work, we study the post-discharge mortality prediction of patients by deriving insights from the unstructured clinical descriptions (UCDs). The study by Churpek et al. (2016) concluded that in clinical practice, simpler models are most commonly deployed, which are easy to interpret by physicians as well as regulators. Hence, our work will be focused on building simple as well as interpretable models which can be used for building clinical decision support tools that flag at-risk patients.

The key contribution of our work is we have built a simple BOW model for 7/30/180/365 day mortality prediction just based on clinical notes. The performance thus achieved (0.92 AUC for 30 day mortality) is good given its simplicity. We also explored inbuilt library based syntactic sentiment analysis for doing preliminary survival analysis.

The rest of our paper is organised as follows — Section: 2 describes the dataset used, patient cohort selection and routine data preprocessing methodology. Section: 3 describes various models used in the experiment. Finally, Section: 4 presents the evaluation metric we have chosen, their interpretation and discussions around it.

2 Methodology & Dataset

2.1 Dataset

We used MIMIC-III v1.4 Johnson et al. (2016), which is publicly available de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset contains information on patient demographics, vital sign measurements, lab test results, clinical procedures, medications, caregiver notes, imaging reports, and mortality labels.

2.2 Patient Cohort Selection

We retrieved set of adult patients ($\geq$ 18 years) with at least one associated note from the MIMIC-III database. We specifically extracted `Nursing Progress Note` from the dataset. Check the Figure 1 to understand number of notes for each category in whole dataset, we chose Nursing as it is more useful for our purpose as well as it has most number of records.

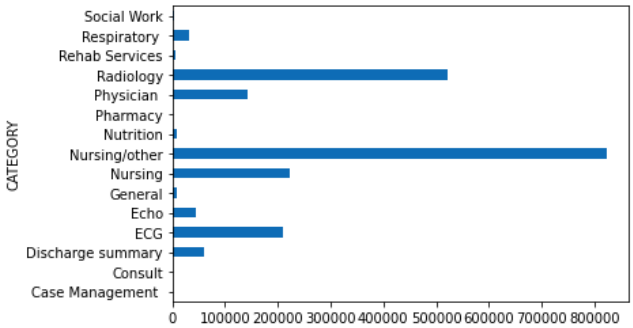


Figure 1: Numbers of Notes Per Category

2.3 Data Preparation

We used google bigquery and its SQL dialect for our patient cohort selection. The mean number of notes per hospital admission is 32 (median: 15). We concatenated all notes of a particular patient with the same hospital admission id. A such concatenated sample note is shown in Figure 2.

```
patient transferred from hospital ward name 289 11 with change in mental status
and n changes on ct scan nlast name problem pneumocephalus n assessment n patient
lethargic but has increased wakefulness as shiftprogresses n follows commands
pupils pearla orientated to self only heart rate n sinus rhythm with lots ofectopy
systolic b p 90 130 ns over 50 ns foley n patent draining clear yellow urine
lumbar drain clamped siteintact n full strength all extremities n action n
echocardiogram done by fellow npo n response n plan n lastname problem
pneumocephalus n assessment n early am placed on open face mask for humidification
wfi02 100 and n 15! abg wnl n weak cough rhonchi in upper lobes minimal secretions
dry mucosa n continuedwith purposeful mvt but nonverbal no obeying of commands n
this afternoon pt presenting with a worseningrespiratory picture hr n 120 ns 130
ns rr 30 40 ns labored with abdominal breathing agitated n sbp 140 ns160 n action
n placed back on bipap n mso4 0 3mg x2 for dyspnea ativan 0 25mg for agitation x3
n albuminas ordered for low u o and tachycardia n lopressor dose increased to 10mg
hr q6 n seroquel SOmg to rectifysleep schedule per neuro surgery n md last name
titles 8721 bedside for sustained hr 120 ns 130 ns and rr30 40 fellow n md doctor
last name 8817 and doctor last name 3965 made aware of developing situation
nfamily 4 daughters and wife made aware of option fro bronch bedside n and
intubation n response n familydoes not want intubation or invasive treatment n
last p02 80 on bipap n cmo with mso4 git when familyarrives n plan n 05 mg of mso4
hourly until pt is officially cmo md doctor last name 8721 made aware n
ofcontinuing hr in 130 ns and rr 30 n non invasive ventilation n pastoral care
bedside n family believes pt iscurrently comfortable and awaiting family before n
mso4 gtt n name8 md 883 md doctor last name 8721when is ready to be cmo n tracheal
tear n assessment n intubated on mmv d t periods of apnea n suctionedfor bloody
secretions n location un 1083 j collar on n sbp pressor dependent n urine output
borderline nappears very fluid overloaded with generalized anasarca n no contact
with son overnight n action n name nicultures obtained from left femoral line n
suctioning minimized n weaned levophed as tolerated at beginningof shift with n
small wean of levo sbp down to 70 milrinone stopped per sicu fellow n dt no
significant changed
```

Figure 2: Sample Concatenated Note

We also created an attribute 'surv_day', which is essentially the number of days the patient has survived after discharge from the hospital. For patients whose clinical file didn't have a death date (i.e., those who are still alive) were marked as NaN. We used this 'surv_day' attribute to build some new binary labels — 'one week mortality' (if surv_day $<$7 then 1 else 0), 'thirty day mortality', 'one eighty mortality', and finally a 'one year mortality' label.

Attribute Label	**Frequency (Prevalence)**
one_week_mortality	258
thirty_day_mortality	736
one_eighty_mortality	253
one_year_mortality	653

Table 1: Prevalence in Survival Class

Patients were randomly split into train (70%), validation (15%), and test (15%) sets. Since we have a limited number of True labels for any class, for example, consider thirty_day_mortality has train prevalence (Number of True in mortality label) is 11%. It has a class imbalance, to deal we this we did a minority class oversampling.

We removed dates, symbols, and punctuation to prevent overfitting; lowercased all the text using regular expressions. To keep things simple, we tokenized our notes and then used the `CountVectorizer` on it. We did not use any embeddings as it may contain some inherent bias. Whole data pre-processing workflow is shown in Figure 3.

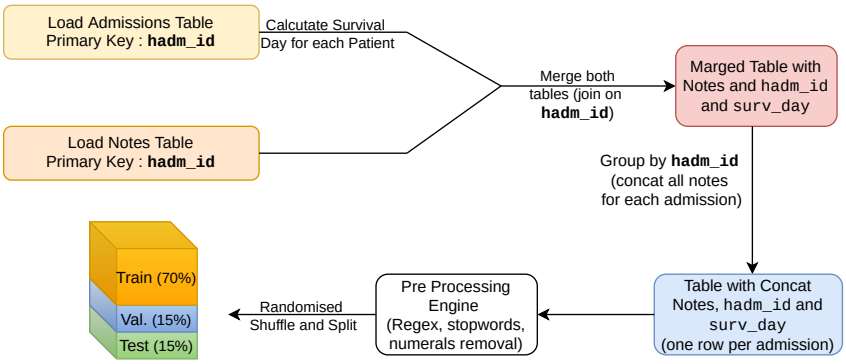


Figure 3: Workflow of data preparation

3 Model Description & Motive

Although recent work by Rajkomar et al. (2018) has achieved 0.94 AUROC for in-hospital mortality; however, they have used sophisticated deep learning techniques. As already explained earlier in motivation, we are keen on building simple as well as an interpretable model for mortality prediction. Hence, we will use simple models such as Logistic Regression, Trees, and Support Vector Machines.

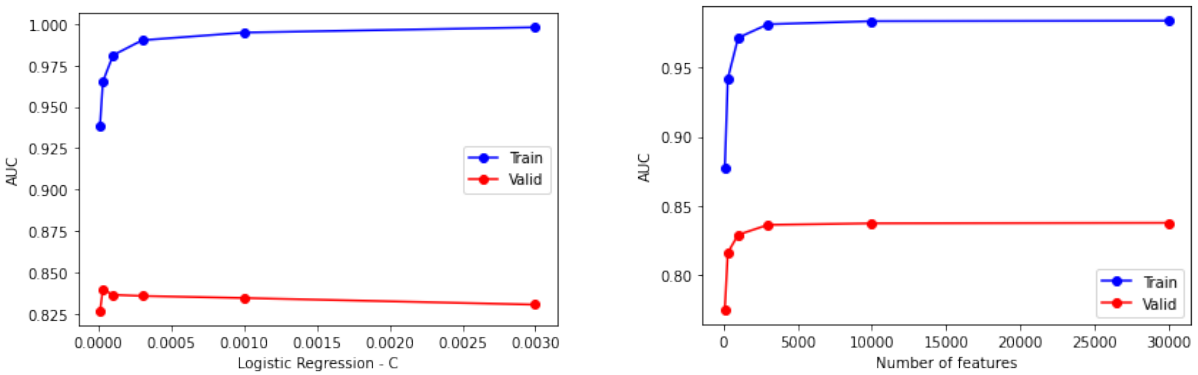
Since we are using bag of words approach with logistic regression for model building we can plot the most important features for each **True** and **False** survival class. This approach helped us in finding out few more stop words that our model is learning which are any-ways redundant.

True = [`ml, should, time, it, been, cm, cc, dr`]

False = [`po, off, c, i, level, to, s, at`]

We added them in our new stop words list subsequently removing them from our vocabulary. We experimented with various hyper-parameters for all models using standard methodology and used ones, which provided best performance. These curves are shown in Figure 4. We also plotted the learning curves to find out that addition of more data is not likely to improve the performance of our simple model.

Sentiment Analysis Following the idea of Waudby-Smith et al. (2018), we tried to see whether our concatenated notes have any sentiment which can be used as a proxy for mortality. Since sentiment analysis is inherently a supervised learning problem but as we don't have any labelled data of true sentiments we have resorted to using syntactic approach and used Textblob NLP Library. Using `v0.16.0` of this library we extracted two syntactic features from the text – polarity and subjectivity. The Textblob algorithm tokenizes the text, performs POS tagging, and uses a lexicon to give a sentiment polarity and subjectivity score. We used these features with SAPSII score (which are already provided in MIMIC dataset) to do the survival analysis of the patients. We partitioned the polarity and subjectivity into



(a) C Value

(b) Number of Features

Figure 4: Choice of Hyperparamter for logistic regression

4 quartiles and plotted the Kaplan Meier curve Kaplan and Meier (1958). The survival analysis was performed in R (v 3.6.3) using `survminer`. First for a preliminary analysis only subjectivity (as obtained from nursing notes – one of the dimension of sentiment) was used to plot the survival curve with $p < 0.0001$. The results are shown in Figure 5. The results for Kaplan Meier curve (plotted using both

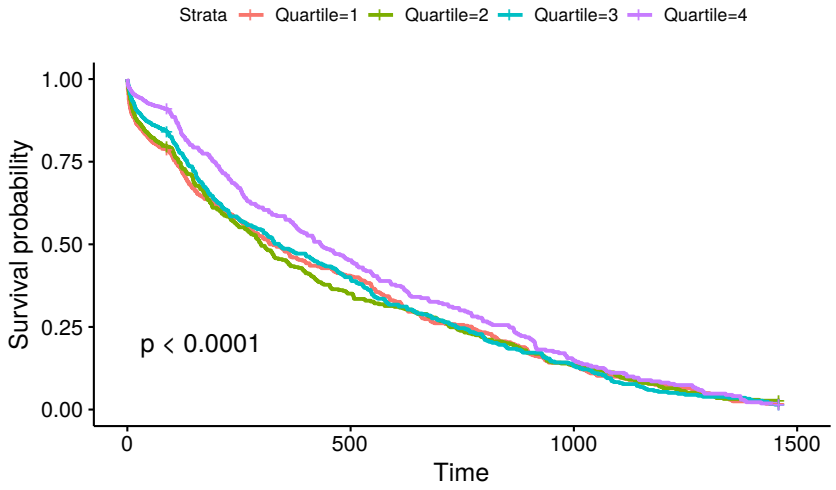


Figure 5: Survival Curves using Only Subjectivity and SAPSII Score

subjectivity, polarity and SAPSII score) is reproduced in Figure 7.

4 Results and Analysis

This section assess the performance of our model. Specifically in next subsection 4.1 we define the evaluation metric and subsequent subsection 4.2 reports the results.

4.1 Evaluation Metrics

- We have used Area under the receiver operator curve (**AUC**) to judge the performance of the model. This metric is consistent with other studies published in the domain Duda et al. (2000).
- Apart from AUC, we reported **Accuracy** as well. Accuracy is defined as $\frac{\sum \text{TP} + \sum \text{TN}}{\sum \text{Total Population}}$.
- We also reported **Specificity**, which is defined as

$$\frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

4.2 Results

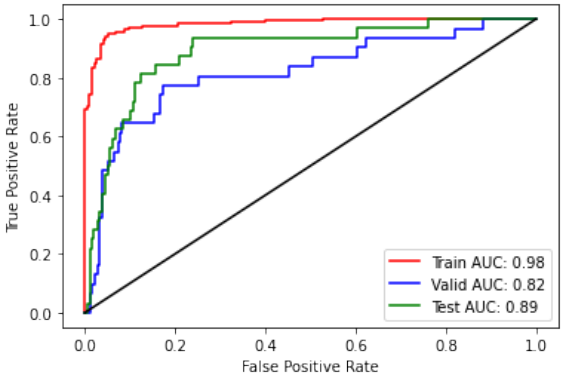
Model →	Logistic Regression			Decision Tree		
Metric →	AUC	Accuracy	Specificity	AUC	Accuracy	Specificity
7 day Mortality	0.891	0.766	0.762	0.743	0.765	0.767
30 day Mortality	0.884	0.855	0.862	0.809	0.810	0.811
6 month Mortality	0.641	0.599	0.597	0.554	0.662	0.671
1 year Mortality	**0.607**	0.569	0.567	0.550	0.542	0.540
Model →	XG Boost			SVM		
Metric →	AUC	Accuracy	Specificity	AUC	Accuracy	Specificity
7 day Mortality	**0.899**	0.834	0.834	0.739	0.679	0.68
30 day Mortality	**0.926**	0.906	0.915	0.665	0.705	0.744
6 month Mortality	**0.678**	0.640	0.640	0.569	0.600	0.609
1 year Mortality	0.604	0.597	0.602	0.500	0.515	0.536

Table 2: Evaluation Metric for Various Models

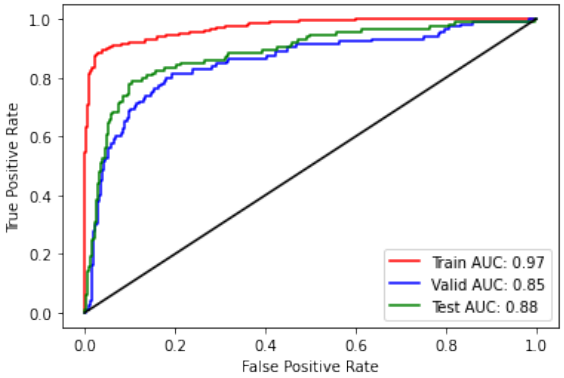
Results from our experiment are presented in Table 2. Among all the models experimented Logistic Regression and XG Boost performed better compared to others. The best AUC was achieved for 30 day mortality prediction and performance drops as we aim for 6 month or 1 year long mortality prediction. The AUC plots for various survival class using simple Logistic Regression model is shown is Figure 6.

We hypothesize that nursing notes have opinions and other latent features which are predictive of near mortality viz. one week or one month. But as we try to predict the longer period mortality such as six month or one year it becomes increasingly difficult.

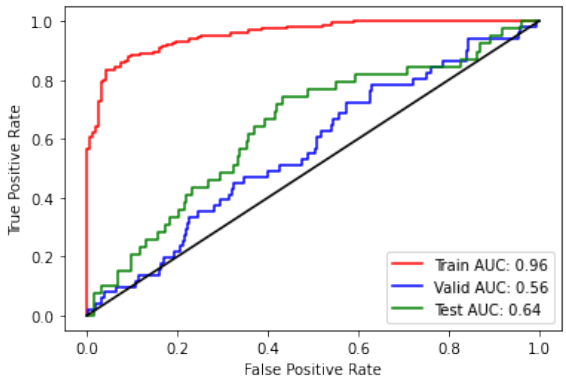
We have plotted the KM curve (in Figure 7) which has polarity and subjectivity quartiles somewhat separated. Top two quartiles are well separated but bottom two have some overlap. Waudby-Smith et al. (2018) in their work have statistically shown that quartiles are nicely separated. In our work may be concatenating the notes and taking all the notes including (discharge notes too) might be a reason for overlap, it needs statistical investigation.



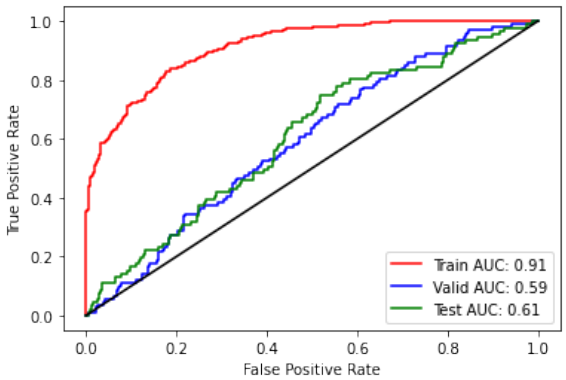
(a) 7 day mortality



(b) 30 day mortality



(c) 180 day mortality



(d) 365 day mortality

Figure 6: AUC Curve for 7/30/180/365 day mortality

5 Discussion & Future Directions

While our model does not beat the state of the art models for mortality prediction from clinical notes; however, it compares favourably with most given its simplicity. It clearly shows that unstructured text in the EHR notes contains meaningful information that can be mined for building intelligent medical decision support systems. Our data is from a single hospital, so patients have somewhat similar demography as well as nurse, and physicians are the same; their way of recording the opinion is somewhat similar. If we could use the data from multiple sources, we can genuinely validate our claim.

We plan to incorporate other structural data (which is mostly time series such as patient measurements, etc.) in our model; and build a simple multimodal fusion model that can leverage both structural as well as unstructured free form text. The sentiment of the notes can also be incorporated into the model, as already shown above and empirical work by Waudby-Smith et al. (2018) support this too, that sentiment and opinion of nurse/physician have mortality prediction property. On these lines, we can improve our model's utility.

Prediction of mortality remains a complex problem, however by incorporating more information extracted from unstructured data, like nursing notes it might improve the overall performance.

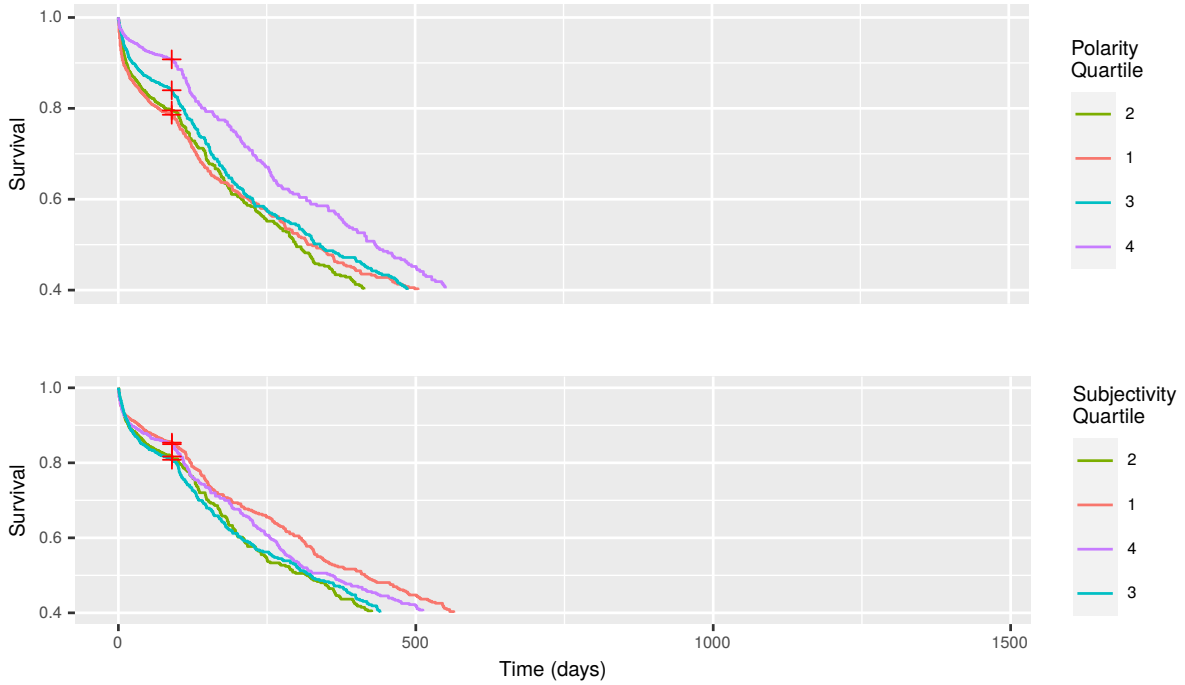


Figure 7: Kaplan Meier Curve

Software and Data

Code is available in github https://github.com/vntkumar8/clinical-notes-mining.
The data is available publicly (to be downloaded) from physionet website https://physionet.org/content/mimiciii/1.4/. Due to confidentiality agreement signed between dataset provider, we can not share the data.

References

Matthew M Churpek, Trevor C Yuen, Christopher Winslow, David O Meltzer, Michael W Kattan, and Dana P Edelson. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine*, 44(2):368, 2016.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. Comparing rule-based and deep learning models for patient phenotyping. *arXiv preprint arXiv:1703.08705*, 2017.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Mark L Metersky, Grant Waterer, Wato Nsa, and Dale W Bratzler. Predictors of in-hospital vs postdischarge mortality in pneumonia. *Chest*, 142(2):476–481, 2012.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

Bonggun Shin, Julien Hogan, Andrew B Adams, Raymond J Lynch, Rachel E Patzer, and Jinho D Choi. Multimodal ensemble approach to incorporate various types of clinical notes for predicting readmission. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE, 2019.

Michael Staff. Can data extraction from general practitioners' electronic records be used to predict clinical outcomes for patients with type 2 diabetes? *Journal of Innovation in Health Informatics*, 20 (2):95–102, 2013.

Ian E. R. Waudby-Smith, Nam Tran, Joel A. Dubin, and Joon Lee. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PLOS ONE*, 13(6):1–11, 06 2018. doi: 10.1371/journal.pone.0198687. URL https://doi.org/10.1371/journal.pone.0198687.