DEVELOPING SOFTWARE SOLUTIONS FOR EHR ANALYTICS

Shayantan Banerjee

Team: maric2018

Affiliation: Indian Institute of Technology, Madras

ELECTRONIC HEALTH RECORDS

- A structured approach to storing digitized medical documents
- 99% of US hospitals have some sort of EHR system in place
- Huge predictive power
- Features can be extracted from clinician's notes to train analytical models
- These are usually available publicly and are de-identified to protect patient privacy

SmartEHR

A predictive analytics approach to analyze free text clinical notes

Enter Your Text Below

Submit Clear Reset

MIMIC III DATASET

- We considered MIMIC III (Medical Information Mart for Intensive Care III) free hospital database. This database contains de-identified data from over 50,000 patients who were admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012. In order to get access to the data for this project, you will need to request access at this link (https://mimic.physionet.org/gettingstarted/access/).
- These are free text clinical notes
- These are highly unstructured and machine unintelligible

STRUCTURE

- 'Admission Date: [**2106-4-6**] Discharge Date: [**2106-4-15**]\n\nDate of Birth: [**2038-4-1**] Sex: M\n\nService: MEDICINE\n\nAllergies:\nPatient recorded as having No Known Allergies to Drugs\n\nAttending:[**First Name3 (LF) 1990**]\nChief Complaint:\nFever, hypotension\n\nMajor Surgical or Invasive Procedure:\nBedside debridement of ulcerations by plastic surgery team\n\n\nHistory of Present Illness:\n68M with h/o t4 paraplegia x 2yrs, felt [**3-13**] "inflammatory spinal\ndisease", with a chronic indwelling foley, sacral decubitus\nulcers, presents to [**Hospital1 18**] from rehab after RN noted 1d of fever\n(tmax 101.8). [**Name8 (MD) **] RN caring for pt at rehab, pt noted some mild\nabdominal discomfort (chronic), but otherwise denied any recent\nsymptoms of cough, n/v, constipation, rash. Pt has been having\nchronic diarrhea (x3/day, x2-3/night) for past 1yr, etiology\nunclear. [**Name2 (NI) 227**] persistent fevers x24hrs, pt was brought to\n[**Hospital1 18**] ED. [**Name8 (MD) **] RN BP prior to leaving rehab was 100/72.\n.\nPer pt, he notes chronic abdominal pain, "always there",\ndiffuse, sharp, sometimes awakening him from sleep, no relation\nto food or BMs. somewhat worse over the preceding 4 months, but\nactually improving over the past few days. At present, he\nstates his pain has completely resolved. ROS otherwise\nsignificant for +orthopnea, pt also notes nonproductive cough x\n3 weeks, no flu sx (body aches, congestion, sore throat). Pt\ndenies flut shot or pneumovax. +sick contacts (lives in [**Hospital 100**]\nRehab).\n.\nUpon arrival in ED VS=100.4 100 87/51 12 95%RA. UA was c/w\nUTI, pt was started on vanco and zosyn, UCx and BCx sent.\nsacral ulcers felt to be stage 4, no evidence of superinfection.\n BP initially responded to 3L IVF (99/53), however after 3rd\nlitre, BP down to 85/40, pt therefore received RIJ TLC, and\npossibly an additional 1L IVF bolus, afterwhich BP improved to\n115/70. Pt was asymptomatic, mentating throughout without\nspecific complaints.\n.\nPt also noted moderate abdominal tenderness. CT ABD done which\nshowed no acute processes. CXR unremarkable, EKG unremarkable\n(old Q in III, ?mild ST changes V1).\n.\nPt admitted to ICU for further monitoring given hypotension.\n.\n\nPast Medical History:\n1. Inflammatory disease of the spinal cord of uncertain\netiology. MRA [**10-16**] negative for vascular malformation. Initial\nCSF analysis showed elevated protein (82) without oligoclonal\nbands. NMO blood titer negative, RPR negative, Lyme serology\nnegative, [**Doctor First Name **] negative, Ro and La negative, ACE level normal,\nneuromyelitis IgG negative, ESR 70, CRP 66.8. Ultimately\ntreated with broad spectrum antibiotics, corticosteroids (two\nweeks of Solu-Medrol followed by a prednisone taper), and 5 days\nof mannitol without improvement. He is followed by neurology\nfor a dense paraplegia (T4) with neuropathic pain, restrictive\nshoulder arthropathy, and a neurogenic bladder requiring a\nchronic indwelling foley.\n2. Chronic sacral decubitus ulcer, previously treated with a VAC\ndressing\n3. Multiple UTI (including Pseudomonas)\n4. Pulmonary embolus [**11-15**] s/p IVC filter placement\n5. Asthma\n6. Two-vessel coronary artery disease s/p CABG 4-5 years ago\n7. Systolic CHF (EF 25-30% on [**2-15**] TTE)\n8. Repaired liver laceration\n9. Chronic back pain\n10. Vitiligo\n11. Feeding tube\n12. Depression\n13. MRSA from sacral swab and sputum\n14. Prior transient episodes of leg paralysis\n15. Right frontal lobe brain lesion biopsied [**11-15**] and c/w\ngliosis; resolved on repeat imaging\n16. Abnormal visual evoked potentials\n\nSocial History:\nHe moved here from [**Country 3594**] (after living in many different\ncountries) in the [**2068**]. He is retired from a job in the\nmaritime industry. Divorced 24 years ago. Three children.\nQuit smoking [**2076**]. Quit drinking [**2080**]. No history of illicit\ndrug use or abuse.\n\n\nFamily History:\nNo stroke, aneurysm, no seizure, no AAA.\n\nPhysical Exam:\nVS: 96.6 85 105/66 15 100%2L\nGen: Well appearing male in NAD lying in bed.\nHEENT: JVD <6-8cm, MMM, lips slightly pale.\nChest: CTA bilaterally, no w/r/r.\nCV: RRR, physiologic splitting S2, no r/g. 3/6 SEM @ LSB.\nAbd: Soft, nontender to deep palpation in all four quadrants,\ndistended, tympanic (?gas), negative murphys sign, well-healed\nmidline g-tube scar.\nExtremities: Warm, well perfused, no C/C. [**2-10**]+ edema bilaterally\nto knees.\nSkin: Vitiligo on hands. Large round 6x4 cm diameter pressure\ndecubitus ulcer on sacrum and 4x3cm decub ulcer on left ischial\ntuberosity. Appears clean with granulation tissue in center, no\ns/sx of infection. no purulent drainage.\nNeuro: CN grossly intact. A&O x 3, pleasantly conversant.\n\n\nPertinent Results:\n[**2106-4-5**] 11:50PM BLOOD WBC-9.08 RBC-4.37* Hgb-11.2* Hct-34.9*\nMCV-80* MCH-25.6* MCHC-32.0 RDW-15.1\n[**2106-4-8**] 04:47AM BLOOD WBC-6.7 RBC-3.49* Hgb-8.9* Hct-28.5*\nMCV-82 MCH-25.6* MCHC-31.4 RDW-14.9\n[**2106-4-5**] 11:50PM BLOOD Glucose-125* UreaN-11 Creat-0.5 Na-137\nK-4.0 Cl-101 HCO3-27 AnGap-13\n[**2106-4-8**] 04:47AM BLOOD Glucose-109* UreaN-5* Creat-0.4* Na-139\nK-3.7 Cl-110* HCO3-23 AnGap-10\n[**2106-4-6**] 10:27PM BLOOD CK-MB-5 cTropnT-0.08*\n[**2106-4-6**] 08:11AM BLOOD cTropnT-0.08*\n[**2106-4-5**] 11:50PM BLOOD CK-MB-NotDone cTropnT-0.09*\n[**2106-4-8**] 04:47AM BLOOD Calcium-8.2* Phos-3.0 Mg-2.0\n[**2106-4-6**] 12:05PM BLOOD Cortsol-15.3\n[**2106-4-6**] 12:05PM BLOOD CRP-122.0*\n[**2106-4-6**] 01:45PM BLOOD Lactate-1.4\n[**2106-4-6**] 12:00PM BLOOD Lactate-0.7\n[**2106-4-6**] 12:02AM BLOOD Lactate-1.7\n\nCT ABD/Pelv [**2106-4-6**]:\n1. Severe sacral and right ischial tuberosity decubitus ulcers.\n2. No acute intra-abdominal inflammatory process.\n3. Cholelithiasis.\n\nCXR [**4-6**] Bedside frontal chest radiograph is compared to\n[**2106-1-2**] and demonstrate clear lungs, normal pulmonary\nvasculature, and no evidence for pleural effusions. The heart\nand mediastinal contours, remarkable for tortuous aorta, are\nstable. This patient is status post median sternotomy.\n\nIMPRESSION: No acute cardiopulmonary process.\n\nEKGs: NSR, essentially unchanged from prior tracings\n\nWBC scan;\n\nIMPRESSION: 1. Unchanged appearance of residual .

  - In [ ]:

OBJECTIVES

- Extract predictive features from free text clinical notes **only** and perform the following tasks:
- Predict what are the chances of a patient getting re-admitted to the hospital within 30 days of discharge **(Hospital)**
- Predict diagnosis codes (ICD9) using word vectors as features **(Doctors/caregivers)**
- Predict the top 5 specific identifiers (keywords) for each diagnosis (Keyword summarization for quick review) **(Doctors/caregivers)**
- **Combine all these into a web-app using the Python Flask framework**

(1) PREDICT 30-DAY READMISSION CHANCES

- Only unstructured text data used
- **Only discharge notes were considered**
- Regex used to filter out unwanted characters
- Stop words identified and removed
- Unbalanced dataset
- Subsampling of majority class done to restore balance
- 33-66 split
- CountVectorizer() used to encode documents
- The created vocabulary was used to encode new documents
- A logistic regression classifier was used

Challenges:

1. Highly imbalanced data (Majority class subsampling)
2. Unstructured text demanded a lot of cleaning (most regex based)
3. Manual validation of the keywords very specific to the model

RESULTS

5-fold cross validation was performed and the fitted model was used for predictions

Train AUC: 0.748

Test AUC: 0.755

Table 4

Comparison of the performance of our models with that of LACE, assuming a 25% intervention rate.

Model*	# Features	Precision	Recall	AUC	Training time**	Evaluation time**
2-layer neural network	1667	**24%**	60%	**0.78**	2650 sec	154 sec
2-layer neural network	500	22%	**61%**	0.77	396	31
2-layer neural network	100	22%	58%	0.76	169	14
Random forest	100	23%	57%	0.77	669	43
Logistic regression	1667	17%	41%	0.66	60	4
Logistic regression	100	21%	52%	0.72	17	0.1
LACE	4	21%	49%	0.72***	0	0.2

Ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5510858/

MANUAL VALIDATION

Below are some of the top common terms that explained 30-day
readmission and their literature reference

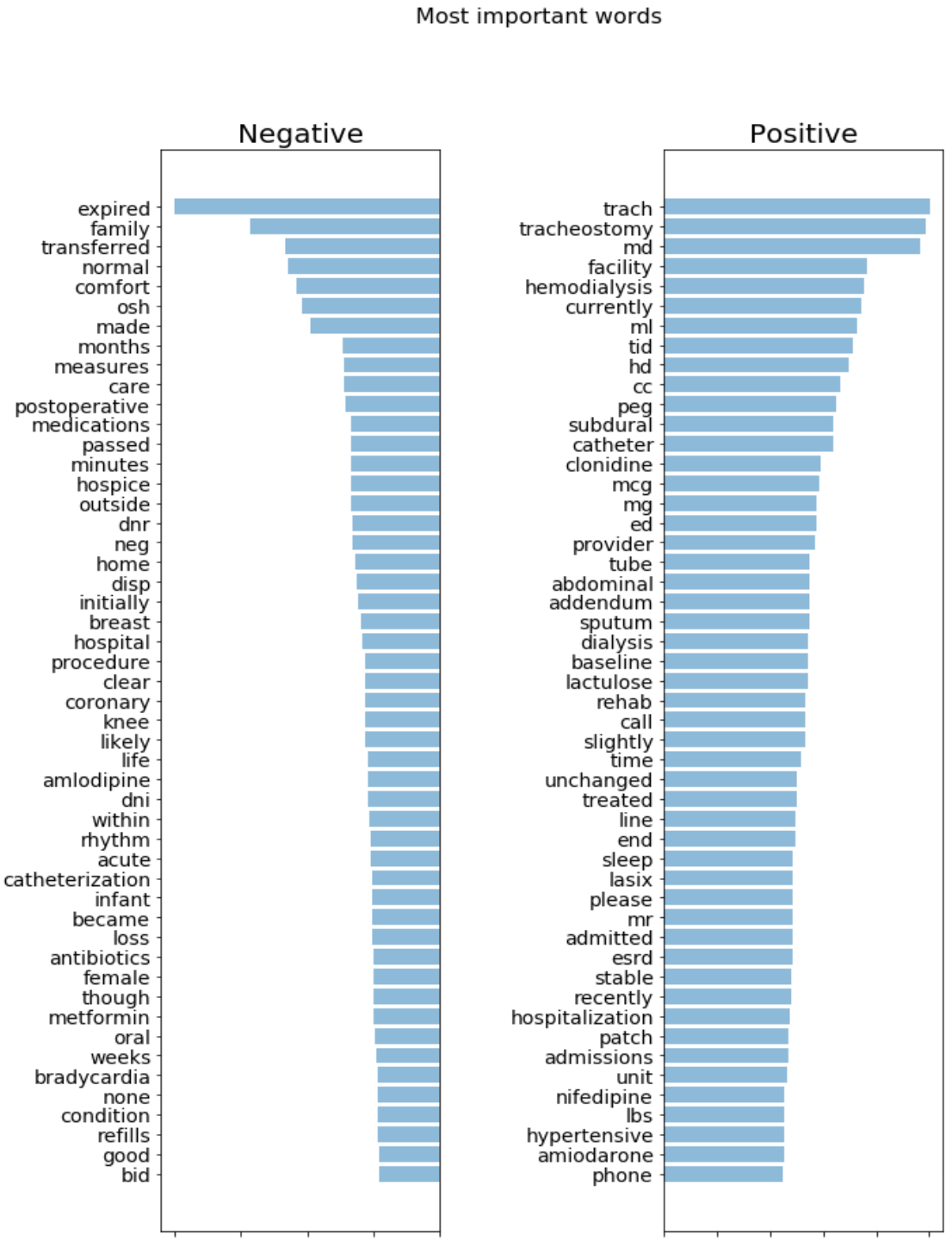
Trach,tracheostomy:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5510858/

Hemodyalisis,ESRD:
https://www.medpagetoday.com/nephrology/esrd/68196

Peg: (percutaneous endoscopic gastrostomy)
https://www.ncbi.nlm.nih.gov/pubmed/27423366

Subdural:
https://www.ncbi.nlm.nih.gov/pubmed/29271714

Figure on the right shows the top word features that explains the model (Positive: Will get admitted within 30 days)
Negative: Not



(2) PREDICT TOP 10 DIAGNOSES CODES

- Top 30 diagnoses were considered as data
- Corresponding TEXT data was extracted
- Only discharge notes were considered
- Same patient --> Same hospital id --> Multiple diagnoses!
- This confounds the problem further
- Makes it a multilabel classification (harder to capture textual features as compared to binary classification)

In [75]:
```
data_cleaned_key.head(10)
```

Out[75]:

	SUBJECT_ID	HADM_ID	ICD9_CODE	CATEGORY	TEXT
0	109	172335	40301	Discharge summary	admission date discharge date date of birth se...
1	109	172335	486	Discharge summary	admission date discharge date date of birth se...
2	109	172335	58281	Discharge summary	admission date discharge date date of birth se...
3	109	172335	5855	Discharge summary	admission date discharge date date of birth se...
4	109	172335	4254	Discharge summary	admission date discharge date date of birth se...
5	109	172335	2762	Discharge summary	admission date discharge date date of birth se...
6	109	172335	7100	Discharge summary	admission date discharge date date of birth se...
7	109	172335	2767	Discharge summary	admission date discharge date date of birth se...
8	109	172335	7243	Discharge summary	admission date discharge date date of birth se...
9	109	172335	45829	Discharge summary	admission date discharge date date of birth se...

Challenges

1. Multilabel classification

2. Some discharge notes had addendum attached to it. Excluding that would have been a mistake

3. Several ICD9 codes are now out of use. This created problem during validation

METHODS

- Multilabel classification
- Facebook's FASTTEXT algorithm was used to formulate word representations
- Data for top 30 ICD9 codes were extracted as they represent almost 80% of the data
- AUC: 0.7912

Note:

- FastText is not a great choice for multilabel classification and other tools like StarSpace exists for that purpose

(3) PREDICT IMPORTANT KEYWORDS

- Keywords which are very specific to that document but not to the corpus
- Useful for doctors/nurses for quick review
- If a word appears frequently in a document, it's important. Give the word a high score.
- But if a word appears in many documents, it's not a unique identifier. Give the word a low score.
- The TFIDF scores of words are directly proportional to their frequency within that document but inversely to that corpus
- Domain knowledge required to explain keywords validity

From the above keywords, a doctor can form a rough idea about the most important words that explain the patient's discharge summary. Eg. "gy" is a unit of radiation given to a patient.
"neu" is an onco-gene common in breast cancers
"lvi" stands for lymphovascular invasion
etc

gy	0.333
neu	0.272
nodes	0.242
lvi	0.213
cycles	0.187
infiltrating	0.162
taxol	0.154
cytoxan	0.148
carcinoma	0.143
path	0.138
mastectomy	0.133
er	0.126
breast	0.124
ductal	0.117
mammoglobin	0.109
mm	0.107
letrozole	0.104
cm	0.103
axillary	0.101
positive	0.1

TO SUM UP...

- Efficient tools for processing large scale free-text clinical notes can be developed and deployed
- Tools specific to the Indian population will be possible only if we get digitized patient records as training data
- Further work in terms of model tuning has to be performed
- Deep Learning!
- I intend to make this tool open source such that anyone can download and use it with minimal technical expertise

DEMONSTRATION