Rare disease prediction from clinical notes using LLMs

Archit Patil

Problem Motivation

- Rare diseases are hard to detect.
- Delay in diagnosis: 5-7 years in the U.S.
- Clinical notes have rich signals but are unstructured data.
- Instead of simply extracting medical terms from the notes for model training, context might be important

ICD Codes

- International Classification of Diseases (ICD) is a standardized coding system for diagnoses.
- Numeric codes assigned to diseases, syndromes, symptoms, etc.
- Versions:
  - ICD 9: Used until 2015.
  - ICD 10: Officially used until 2022
  - ICD 11: Official since 2022..but still ICD 10 is used

Orphanet

- The Rare Disease Encyclopedia
- European reference portal for information on rare diseases and orphan drugs.
- >10,000 rare diseases.. more than 7000 genetic diseases
- Provides:
  - Disease Definitions
  - ICD Mappings
  - Prevalance and inheritance patterns

MIMIC-IV Dataset

- Dataset of clinical notes (250,000 patients)
- Source: Beth Israel Deaconess Medical Center, 2008–2019.
- Contains:
  - admissions, patients, transfers.
  - ICD diagnosis codes.
  - ICD procedure codes.
  - Notes: Clinical Notes, Discharge summary, radiology reports, etc

Challenge with using clinical notes

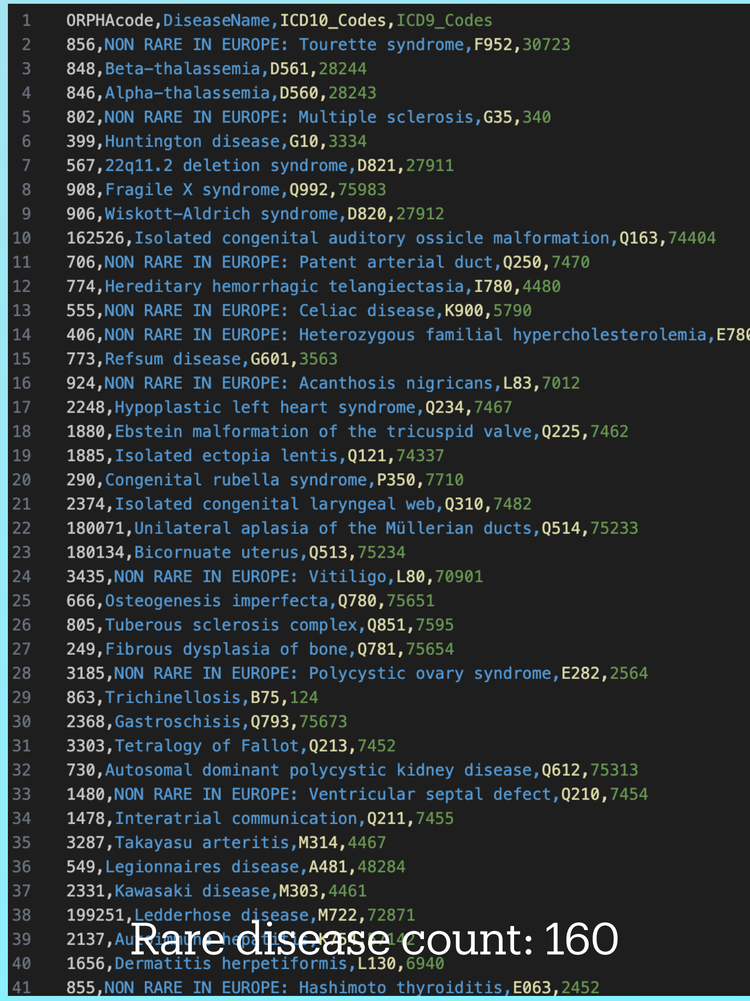
- Long, messy, inconsistent formatting.
- Contains abbreviations, typos, non-standard language.
- Many irrelevant details: need filtering/processing.

Data Preparation

Filtering Rare Diseases

Goal: Identifying rare diseases with exact ICD 9 and 10 mappings to work with

- Orphanet provides ICD-10 and ICD-11 mappings (but no ICD-9).
- We ignore ICD-11 because MIMIC-IV only has ICD-9 and ICD-10.
- Keep only diseases with *exact* ICD-10 matches in Orphanet.
- GEMs = General Equivalence Mappings
- Keep only exact ICD-10 → ICD-9 mappings.
- Handle duplicates



```
ORPHAcode,DiseaseName,ICD10_Codes,ICD9_Codes
856,NON RARE IN EUROPE: Tourette syndrome,F952,30723
848,Beta-thalassemia,D561,28244
846,Alpha-thalassemia,D560,28243
802,NON RARE IN EUROPE: Multiple sclerosis,G35,340
399,Huntington disease,G10,3334
567,22q11.2 deletion syndrome,D821,27911
908,Fragile X syndrome,Q992,75983
906,Wiskott-Aldrich syndrome,D820,27912
162526,Isolated congenital auditory ossicle malformation,Q163,74404
706,NON RARE IN EUROPE: Patent arterial duct,Q250,7470
774,Hereditary hemorrhagic telangiectasia,I780,4480
555,NON RARE IN EUROPE: Celiac disease,K900,5790
406,NON RARE IN EUROPE: Heterozygous familial hypercholesterolemia,E780
773,Refsum disease,G601,3563
924,NON RARE IN EUROPE: Acanthosis nigricans,L83,7012
2248,Hypoplastic left heart syndrome,Q234,7467
1880,Ebstein malformation of the tricuspid valve,Q225,7462
1885,Isolated ectopia lentis,Q121,74337
290,Congenital rubella syndrome,P350,7710
2374,Isolated congenital laryngeal web,Q310,7482
180071,Unilateral aplasia of the Müllerian ducts,Q514,75233
180134,Bicornuate uterus,Q513,75234
3435,NON RARE IN EUROPE: Vitiligo,L80,70901
666,Osteogenesis imperfecta,Q780,75651
805,Tuberous sclerosis complex,Q851,7595
249,Fibrous dysplasia of bone,Q781,75654
3185,NON RARE IN EUROPE: Polycystic ovary syndrome,E282,2564
863,Trichinellosis,B75,124
2368,Gastroschisis,Q793,75673
3303,Tetralogy of Fallot,Q213,7452
730,Autosomal dominant polycystic kidney disease,Q612,75313
1480,NON RARE IN EUROPE: Ventricular septal defect,Q210,7454
1478,Interatrial communication,Q211,7455
3287,Takayasu arteritis,M314,4467
549,Legionnaires disease,A481,48284
2331,Kawasaki disease,M303,4461
199251,Ledderhose disease,M722,72871
2137,Autosomal dominant [illegible]
1656,Dermatitis herpetiformis,L130,6940
855,NON RARE IN EUROPE: Hashimoto thyroiditis,E063,2452
```

Rare disease count: 160

Filtering clinical notes

- Only use notes before the patient is diagnosed with the rare disease
- Model should learn early signals
- Helps prevent data leakage

MIMIC IV Data Extraction

- `diagnoses_icd.csv`: subject ID, admission ID, ICD code/version
- `admissions.csv`: subject ID, admission ID, admit time, other info
- Count occurrences of rare disease ICD codes
- Merge with admission times
- Group by patient and keep first occurrence → first diagnosis date

MIMIC IV Data Extraction

- Add first diagnosis date to each positive patient’s notes
- Negative patients: no diagnosis date
- Keep only notes before first diagnosis date
- Group notes by patient and combine into a single text block

MIMIC IV Data Extraction

- Filter disease list: keep only diseases with >30 positives
- Ensures enough data for model training

```
DiseaseName,ICD9_Code,ICD10_Code,Patient_Count,NumPositivesWithNotes
Beta-thalassemia,28244,D561,66,30
Alpha-thalassemia,28243,D560,60,29
NON RARE IN EUROPE: Celiac disease,5790,K900,818,157
NON RARE IN EUROPE: Vitiligo,70901,L80,119,44
NON RARE IN EUROPE: Polycystic ovary syndrome,2564,E282,903,172
Autosomal dominant polycystic kidney disease,75313,Q612,127,43
Legionnaires disease,48284,A481,115,59
Ledderhose disease,72871,M722,197,92
Autoimmune hepatitis,57142,K754,398,121
NON RARE IN EUROPE: Hashimoto thyroiditis,2452,E063,693,207
Budd-Chiari syndrome,4530,I820,127,60
Guillain-Barré syndrome,3570,G610,272,56
NON RARE IN EUROPE: Idiopathic facial palsy,3510,G510,738,268
Idiopathic eosinophilic pneumonia,5183,J82,105,55
NON RARE IN EUROPE: Hidradenitis suppurativa,70583,L732,259,83
Isolated polycystic liver disease,75162,Q446,89,30
Rare lymphatic malformation,2281,D181,60,31
Von Willebrand disease,2864,D680,219,30
Idiopathic/heritable pulmonary arterial hypertension,4160,I270,344,165
Trigeminal neuralgia,3501,G500,412,139
Arginine vasopressin resistance,5881,N251,88,39
Idiopathic intracranial hypertension,3482,G932,287,81
NON RARE IN EUROPE: Normal pressure hydrocephalus,3315,G912,314,106
Celiac artery compression syndrome,4474,I774,225,111
Pulmonary non-tuberculous mycobacterial infection,310,A310,143,51
Hereditary arginine vasopressin deficiency,2535,E232,280,75
Cholangiocarcinoma,1551,C221,486,205
NON RARE IN EUROPE: Immunoglobulin A deficiency,27901,D802,106,42
Pyoderma gangrenosum,68601,L88,97,46
Secondary polycythemia,2890,D751,400,162
```

Final Disease Count: 30

Control Matching

- Select matched negative patients for each positive
- Matching: age at first visit, gender, race
- Assign pseudo-diagnosis date
- Combine pre-pseudo-diagnosis notes

MIMIC IV Data Extraction

- Filter disease list: keep only diseases with >30 positives
- Ensures enough data for model training

Model Fine Tuning

Target Model and Disease

- Model: ClinicalBERT (BERT trained on clinical notes)
- Target Disease: Nephrogenic Diabetes Insipidus (Arginine Vasopressin Resistance)

    Types: Congenital (genetic), Acquired (Due to external factors)

    Congenital NDI rarity: 1 in 100,000 to 1 in 200,000

Training Setup

- Highly skewed dataset: rare disease prediction
- Pre-split dataset: Training 70%, Validation 10%, Test 20%
- Split maintains positive/control ratio across sets
- Consistent splits used across experiments

Sliding window approach

- ClinicalBERT max context: 512 tokens
- Notes split into chunks: 512 tokens each
- Stride: 256 tokens (overlap)
- Prediction per chunk → average across chunks for final patient prediction

Model Fine Tuning

- ClinicalBERT fine-tuning
- Only the last 4 layers trained (faster and similar/better performance)
- Early stopping applied based on validation loss

Handling Class Imbalance

Loss functions tested:

- Weighted cross-entropy
- Focal loss < final strategy

Oversampling also tested

Learning Rate Strategy

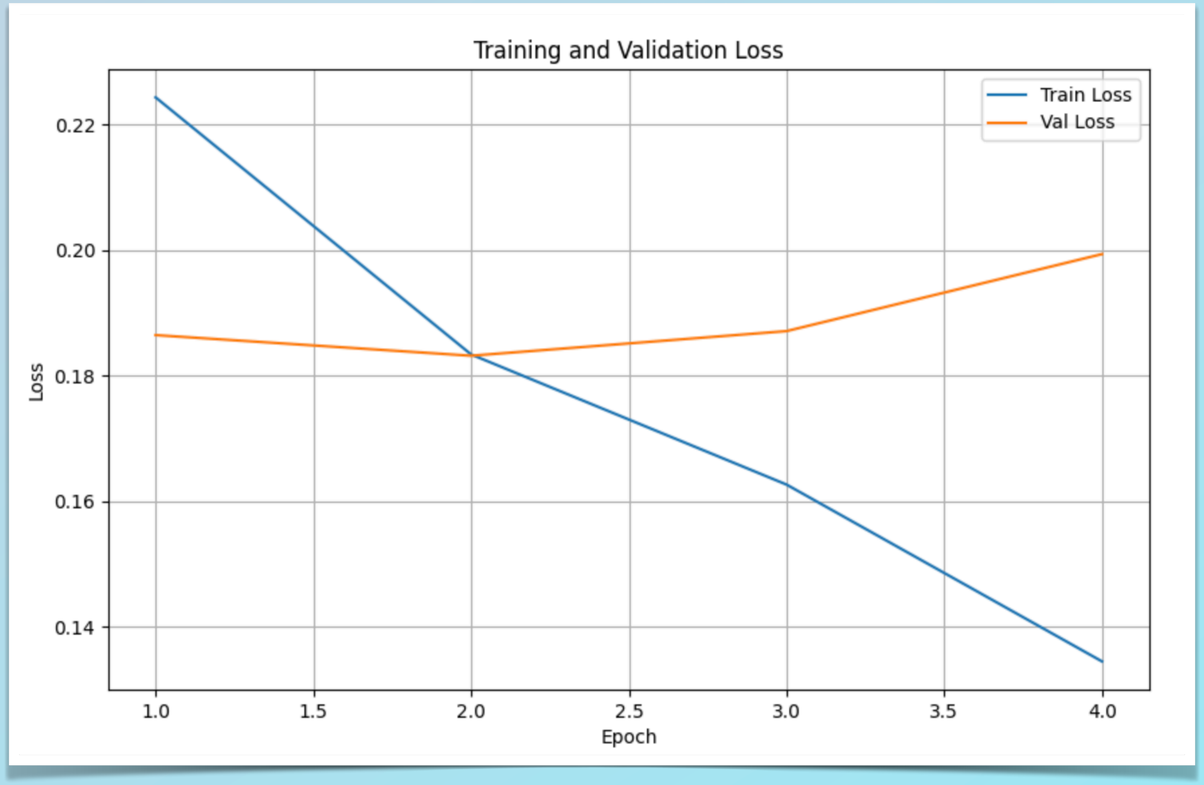
- Learning rate types:
  - Constant
  - Linear decay < final LR scheduler type
  - Step decay
  - Plateau

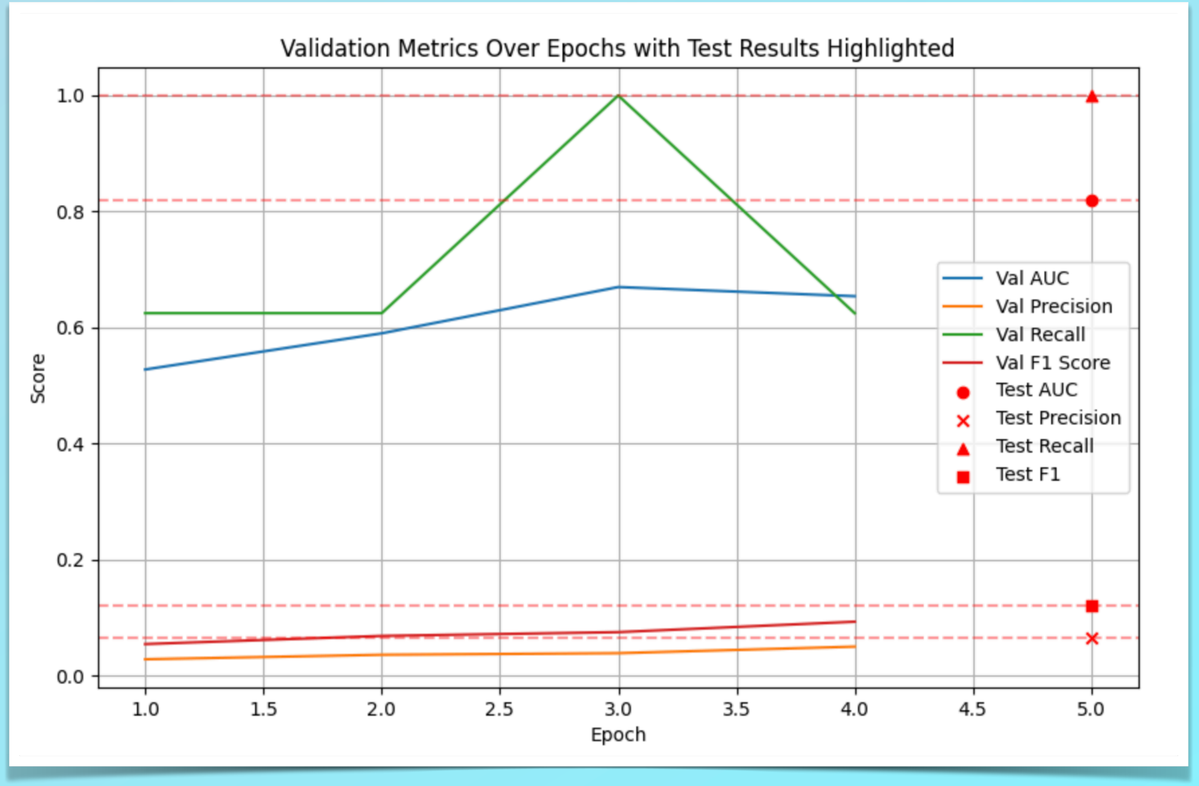
Model Evaluation

Same train/val/test split across different experiments

5-fold cross validation

Results





General AUC: 0.7 to 0.8

5-fold cross validation AUC: 0.78

Comparing models

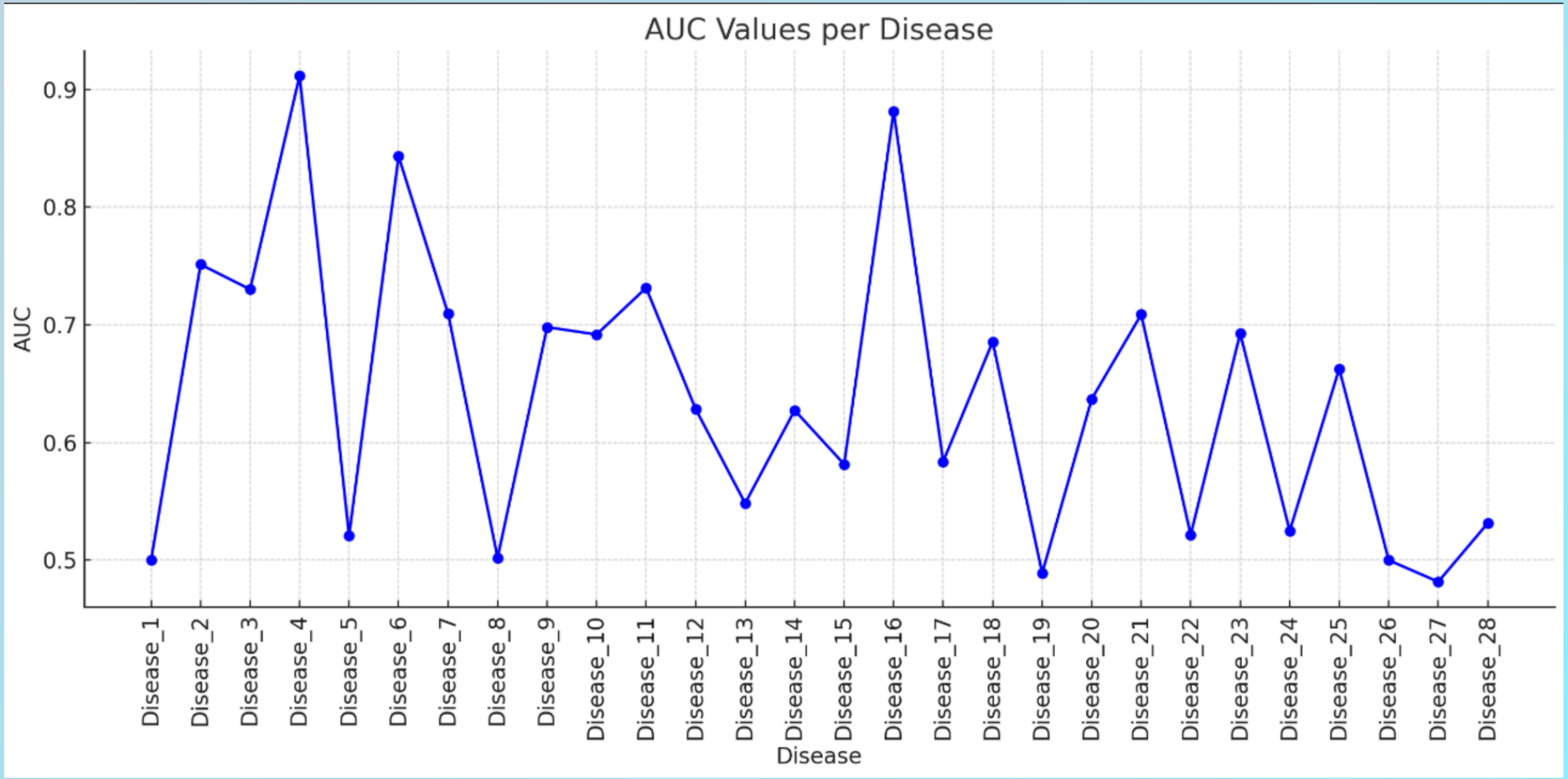
Does context length make a difference? Will longer context model perform better as compared to pre trained BERT models?

Models:

BERT Based models (512 context length): ClinicalBERT, BioBERT, BioLinkBERT-base , PubMedBERT, BioROBERTa, BlueBERT

Long Context models: ModernBERT-Large(8,192 tokens), Longformer (4096 tokens)

Results

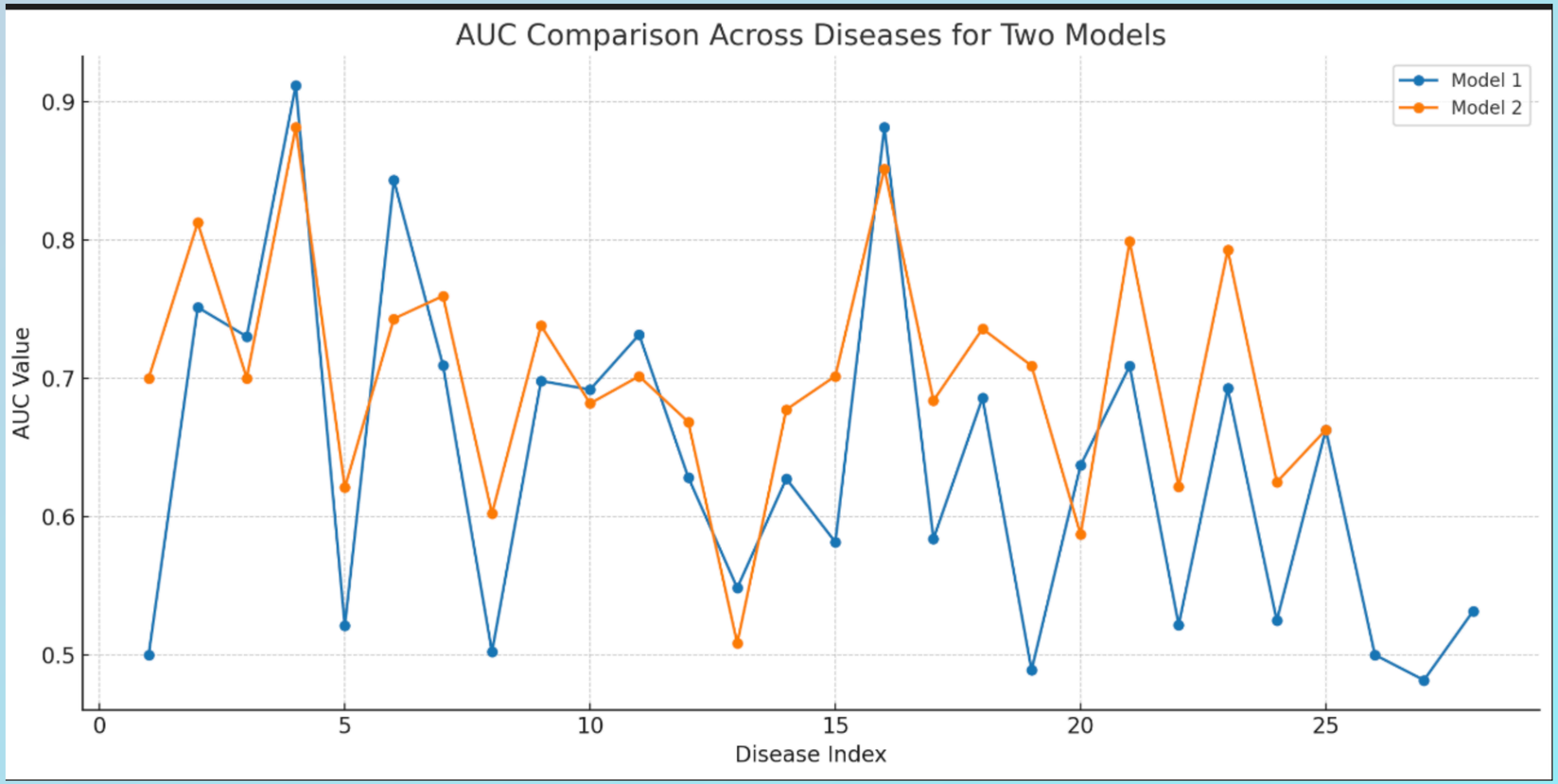


Note: No cross validation done, normal train, val,test splits

BlueBERT-pubmed: trained on pubmed abstracts

Average AUC: 0.67

Results



Model 1: BlueBERT pubmed

Model 2: Clinical BERT

Interpretation

Intepretation

- AUC may not reflect performance on rare positives
- Doctors prefer interpretable results
- Approach 1: Note Highlighter
  - Extract attention values for each token
  - Identify top 5% tokens by attention
  - Highlight corresponding words in the clinical note
  - Highlight intensity $\propto$ attention value

Example

guitarist with four albums and that he makes staggering money he is incredibly grandiose saying when you have a voice like you realize you can have anything you want telling he has 9 **sons** named the first through the **ninth** when you have a voice like you see the value in replicating that at one point **he** tells me im a jillionnaire and adds do you know how many zeroes are in a jillion pause **he** reports being compliant with his medications **he** feels that they keep him on even keel past medical history past medical history irondeficiency anemia **hyperlipidemia** past psychiatric history diagnose **schizoaffective** disorder **bipolar** disorder prior hospitalizations patient reports 5 prior hospitalizations last one years ago at history of assaultive behaviors per **his therapist** he has been violent in the remote past when asked what the most violent thing is that **he** has ever done tells **me** that **he** is a black belt in **karate** and that **he** was in a tournament in and killed a when **he** kicked him 22 times history of **suicide** attempts or selfinjurious behavior denies prior **med** trials unknown per omr has been on **zyprexa** in **the** past therapist at psychiatrist previously dr at now transitioning to dr due to dr patient likes and trusts **his** treaters social history substance abuse history etoh denies recent use tobacco smokes pack per day daily marijuana use history of **crackcocaine** abuse denies recent use social history family history family history mother took when he was younger not sure if she had a formal psychiatric diagnosis no known family history of psychosis or **suicide** physical exam mental status exam arousal awake alert appearance male with long dreads carefully groomed goatee wearing sunglasses so people cant recognize me wearing hospital gown and no pants sits with legs and underwear fully exposed behavior sits in ed gurney looks towards me when talking unable to assess **ec** due to sunglasses no pmrpma but gestures with hands a lot when talking speech normal volum increased prosody increased rate pressured and hyperverbal at times some odd use of language ie i shortrailed **her** income mood good affect grandiose **paranoid** around group home staff somewhat suspicious at times thought process hyperverbal circumstantial extremely grandiose however generally goaldirected answers to questions thought content as per hpi talks about **his** rock music jillions of dollars being upset with group **home** staff judgment and insight poor poor cognitive exam orientation person x place x situation x day x month x **year** x proverbs **the** apple doesnt fall far from **the** tre you have a person like **your** father then your son then **they** have a genetic thing going on loose lips sink ships people like to talk to you and then they spread it around and it ends up in **the** cookery ar then you never have **your** say attentionconcentration focus spells world backwards complex attentionworking memory moyb only able to get to with moyb inattentionneglect not present memory fund c knowledge president obamy current events aware of s **death** comprehension follows simple and multistep commands without difficulty or leftright confusion calculations in 100 x in 135 26 frontalexecutivediffuse visuospatial clock ten past one draws 3dimensional big with clock hands that look about right for 110 physical examination vs bp 14197 hr 110 temp 973 resp 16 o2 sat 100 height weight gen aa **male** with dreadlocks wearing dark sunglasses grooming and hygiene intact heent normocephalic oropharnyx clear perrl eomi cv rrr no mrg pulm ctab abd bs ndnt soft ext no cce skin no **rashes** neurological cn ii **perrla** iii iv vi eom intact v facial sensation intact bilat vii facial mvts symmetric bilat viii finger rub audible bilat ix x palate raise symmetric bilat xi trapezii bilat xii tongue midline motor normal bulk and tone strength in all extremities no pronator drift or asterixis sensory grossly intact to light touch and pinprick bilaterally reflexes 2 biceps triceps brachioradialis and patellar reflexes cerebellar intact finger to nose gait intact abnormal movements none noted pertinent results bmp na 142 k 41 cl 108 co2 27 bun 12 cr 09 glu 92 ag 11 cbc wbc 8 hct 423 plt 226 mcv 7 ua normal serum tox negative urine tox negative **lithium** 0 legal 2 psychiatric on admission the patient was found to be **manic** and psychotic but in good behavioral control it was felt that **his** decompensation was secondary to increased **thc** use recent changes in **his** group home staffing and of difficulty with medication compliance although **his** depakote and **lithium** levels were threapeuti upon admission and **he** noted liking his medications and remained motivated to take them in the hospital per **his** outside treaters although he does have baseline of **hypomania** and fixed delusions they too agreed he was off baseline and **usually** has been unreliable in the past reporting substance abuse during hospitalization the pt was initally restarted on **his** home medications of **lithium** depakote cogentin and propranolol **his** home dosage of prolixin was increased to total 40mg daily with good effect as a result he was calmer and more agreeable to accepting an apology from the group home a meeting was setup with group home staff to address the issues and concerns around admission prior to discharge home at time of discharge **he** was pleasant cooperative and had returned to **his** baseline he was discharged back to **his** group home with outpatient follow up 3 medical for the medical history of hyperlipidemia and **iron** deficiency anemia **he** **was** continued on **his** home medication of ferrous sulfate 325 mg po daily no acute medical issues arose during **his** hospitalization medications on admission depakote er 500 mg am and hs **lithium** eskalith 450 mg bid prolixin 10 mg am 20 mg **hs** 2 mg bid prn cogentin 2 mg bid inderal 20 mg hs ferrous sulfate 325 mg daily discharge medications 1 **lithium** carbonate 450 mg tablet extended release sig one 1 tablet extended release po bid 2 times a day disp60 tablet extended releases refills0 2 benztropine 1 mg tablet **sig** two 2 tablet po bid 2 times a day disp120 tablets refills0 3 propranolol 10 mg tablet sig two 2 tablet po qhs once a day at bedtime disp30 tablets refills0 4 ferrous sulfate 300 mg 60 mg **iron** tablet sig one 1 tablet po daily daily disp30 tablets refills0 5 divalproex mg tablet extended release 24 h sig one 1 tablet extended release 24 hr po daily daily disp30 tablet extended release 24 hrs refills0 6 divalproex mg tablet extended release 24 hr sig two 2 tablet extended release 24 hr po qhs once a day at bedtime disp60 tablet extended release 24 hrs refills0 7 fluphenazine hcl 10 mg tablet sig 25 tablets po qhs once a day at bedtime disp75 tablets refills0 8 fluphenazine hcl 10 mg tablet sig 15 tablets po daily daily disp45 tablets refills0 discharge disposition home discharge diagnosis axis i **schizoaffective** do **bipolar** type axis ii deferred axis iii **iron** def anemia **hypercholesterolemia** axis iv living situation **chronic** **mental** illness discharge condition ambulating independently denies **sihi** mental status exam appearance middle aged aam with long dreadlocks red raincoat casual clothing clean wellgroomed appears younger than stated age behavior sitting calmly in chair good eyecontact to interview no tremor ridgity abnormal movments noted speech reg rate rhythm volume tone mod

Intepretation

Identify top 5% tokens by attention

Map tokens to corresponding words

Create a word list with:

- Word
- Number of occurrences across notes

Example

Arginine vasopressin resistance:

Bipolar, Lithium, Schizophrenia, Hypernatremia, Polyurea, Polydipsia, Diabetes, Mellitus, Dehydration, Medicine, Sex, Surgery,Paranoid, Cancer , Oncologist, Hiv, manic, Atorvastin, Difficulty, doctor

Questions?

Thank you

Acknowledgements: Prof. Vikas Pejaver, Dr. Mohammed Syed, and lab members!