RareBERT: Transformer Architecture for Rare Disease Patient Identification using Administrative Claims

PKS Prakash[1], Srinivas Chilukuri[2], Nikhil Ranade[3], Shankar Viswanathan[4]
ZS Associates, Safina Towers-South Block, 5th Floor, Ali Asker Road, Bengaluru 560052, Karnataka, India[1,3,4]
ZS Associates, 1560 Sherman Ave, Evanston, IL 60201, US[2]
prakash.prakash@zs.com[1], srinivas.chilukuri@zs.com[2], nikhil.ranade@zs.com[3], shankar.viswanathan@zs.com[4]

Abstract

A rare disease is any disease that affects a very small percentage (1 in 1,500) of population. It is estimated that there are nearly 7,000 rare disease affecting 30 million patients in the U. S. alone. Most of the patients suffering from rare diseases experience multiple misdiagnoses and may never be diagnosed correctly. This is largely driven by the low prevalence of the disease that results in a lack of awareness among healthcare providers. There have been efforts from machine learning researchers to develop predictive models to help diagnose patients using healthcare datasets such as electronic health records and administrative claims. Most recently, transformer models have been applied to predict diseases BEHRT, G-BERT and Med-BERT. However, these have been developed specifically for electronic health records (EHR) and have not been designed to address rare disease challenges such as class imbalance, partial longitudinal data capture, and noisy labels. As a result, they deliver poor performance in predicting rare diseases compared with baselines. Besides, EHR datasets are generally confined to the hospital systems using them and do not capture a wider sample of patients thus limiting the availability of sufficient rare disease patients in the dataset. To address these challenges, we introduced an extension of the BERT model tailored for rare disease diagnosis called RareBERT which has been trained on administrative claims datasets. RareBERT extends Med-BERT by including context embedding and temporal reference embedding. Moreover, we introduced a novel adaptive loss function to handle the class imbalance. In this paper, we show our experiments on diagnosing X-Linked Hypophosphatemia (XLH), a genetic rare disease. While RareBERT performs significantly better than the baseline models (79.9% AUPRC versus 30% AUPRC for Med-BERT), owing to the transformer architecture, it also shows its robustness in partial longitudinal data capture caused by poor capture of claims with a drop in performance of only 1.35% AUPRC, compared with 12% for Med-BERT and 33.0% for LSTM and 67.4% for boosting trees based baseline.

Introduction

Identifying the right patients at the right time has always been a primary goal of the life science industry, to improve healthcare services. Individually rare diseases are uncommon; however, there are approximately 8,000 described rare diseases mostly with onset in childhood (Elliott *et al.* 2015 and Zurynski *et al.* 2017) affecting approximately 30 million of US population and 350 million globally worldwide (Colbaugh and Glass 2020). Identifying these patients is a challenging task as patients go undiagnosed or mis-diagnosed for years and move across multiple physicians (Rafi, 2016 and Zurynski Y, *et al.* 2017)

In the health care domain, with an increase in capture of patients' longitudinal history via administrative claims and electronic health records (EHR) enables healthcare industries to understand patient progression or patient identification use cases such as onset of a condition, disease progression or treatment drop-off which in-turn helps to maximize patient care and experience. Researchers are using machine learning based approaches to address patient progression and patient identification related problem. However, owing to the richness of the data; mining this high dimension sequential data is a challenging task.

The current research focuses on rare disease patient identification using administrative claims data. The problem involves many challenges related to data and modelling which includes: 1) Class-imbalance; 2) Unlabeled patients (negative class is not present); 3) Noisy labelled patients; 4) Diagnosis codes for rare condition are not present; and 5) Sensitivity analysis of model due to over-fitting on low positive classes. The current paper focuses on addressing the class imbalance and unlabeled patients challenge. Here, unlabeled

patients are also referred as eligible patients and are ones who have shown presence of comorbidities associated with target disease area such as X-linked hypophosphatemia (XLH) however cannot be classified either as positive or negative class due to missing confirmatory test leading to absence of negative class in dataset. Additionally, impact of noisy scenarios are evaluated on performance of the models.

The current paper proposes a novel RareBERT architecture for feature representation. The proposed architecture extends the Med-BERT architecture to address highly imbalanced class in rare conditions and make improvements in architecture by including context-based embedding, temporal reference embedding and adaptive loss function for faster convergence. Feature representation from RareBERT is integrated with positive unlabeled (PU) learning based architecture also referred as one-class classification for patient identification (Denis *et al.* 2004). The PU learning algorithm trains a classifier on positive and unlabeled data and, estimates the propensity of being a positive class among unlabeled datasets using an adjusted threshold. The efficacy of proposed approach is validated through X-linked hypophosphatemia (XLH) rare condition case study. The experimentation shows significant improvement and robustness of RareBERT with baseline methods.

Rest of the paper is organized as follows: Section 2 presents related prior work; Section 3 describes RareBERT architecture and positive unlabeled (PU)-learning based approach for patient identification; and Section 4 presents the performance of RareBERT using XLH rare condition as a case study, followed by conclusions and proposed next steps in Section 5.

Related Prior Work

Most of the previous work has been focused on solving the following challenges:

1) Lack of specific diagnosis codes to identify patients with the rare disease, thus making the identification harder (no gold standard definition to identify positive patients) (Colbaugh, R *et al.* 2018)
2) Lack of markers to identify true negative patients as a lot of patients are either misdiagnosed or remain undiagnosed for a long time (Yu. K., *et al.* 2019, Garg *et al.* 2016)
3) Low prevalence of positive classes thus leaving a low sample size to learn the patterns from as well as creating a high data imbalance problem (Li W *et al.* 2018, Dai and Hua 2016, Hu *et al.* 2019).

Multiple approaches have been suggested to handle the low positive samples, patients ranging from performing a random under sampling (Dai and Hua 2016) to using Generative Adversarial Networks (GAN) on image representations of patients (Li W *et al.* 2018). For handling the lack of markers challenge Yu. K., *et al.* (2019) suggests using a Sequence Modeling with Generative Adversarial Networks that will help identify potential patients from the unlabeled / undiagnosed patient pool. However, here the assumption is that true negatives can be identified by applying certain rules. In general, such rules are most likely not available for rare diseases specifically.

Recently, BERT (Bidirectional Encoder Representations from Transformers) has been proposed for language representation model for obtaining text embeddings (Vaswani *et al*. 2017). The BERT architecture is proved to be very useful in multiclass and pairwise text classification using transfer learning of pre-trained embeddings. Researchers in life sciences are trying to utilize the BERT architecture on sequential patient journeys to solve problem related to diagnosis code prediction, medication code prediction etc. Architectures such as BEHRT (Li Y., Rao *et al.* 2020), G-BERT (Shang *et a*l. 2019) and Med-BERT (Rasmy *et al.* 2020) are being proposed. BEHRT focused on imputing medical codes within Visit. G-BERT integrated graph neural network (GNN) with BERT architecture with modified masked language to predict medication code prediction using single visit samples which is a limitation. Med-BERT has further extended the architecture to create generalized embedding using bigger vocabulary and used it to optimize two objectives disease prediction and length of stay in hospital. Rasmy *et al.* (2020) has compared different architectures of BEHRT, G-BERT and Med-BERT used for diagnosis and medication tasks. Most of the above architectures are unable to generalize the embeddings when datasets become highly imbalanced. Madabushi *et al.* 2020 has observed imbalance issues in NLP application of propaganda detection and proposed Cost-Sensitive BERT architecture by increasing the weight of incorrect labels.

The feature representation from RareBERT is used for patient identification using the PU learning-based approach. There are a few different PU approaches that have been proposed (Denis *et al.* (2004), Elkan *et al.* (2008), Plessis *et al.* (2015)). Most of these approaches involve isolating a set of so-called true negatives (TNs) from the unlabeled data set. The proposed paper uses biased learning-based approach as proposed by Bekker and Davis (2018) which isolate the negative classes at each iteration.

The next Section presents detailed approach for patient identification using patient administrative claim data.

Approach

The current section presents RareBERT architecture which helps with feature representation. The learned features are passed through PU-learning based semi-supervised classifier to identify patients with high likelihood of having rare condition. The RareBERT architecture helps to deal with

high dimensional sequential data with highly imbalanced classes. The RareBERT architecture enhances Med-BERT architecture by including: 1) *adaptive loss* to balance loss between classification and mask event prediction task; 2) *type embedding* to capture event context such as is it diagnosis, procedure, treatment; and 3) *temporal reference embedding* to capture event position with respect to defined index dates. The next sub-section provides data set-up for RareBERT.

Rare-BERT

The proposed RareBERT set-up uses patient longitudinal claims data. Patient level claims data is right aligned to anchor date and mapped to a Clinical Classification Software (CCS) bucket to roll-up the granular data to a higher resolution as shown in Figure 1.

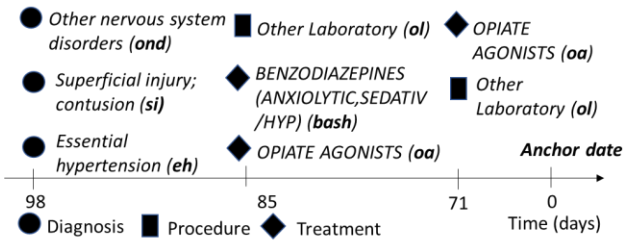

Figure 1: Illustration of patient journey

The patient input sequence is passed through the RareBERT architecture as shown in Fig. 2. The RareBERT determines embeddings from sequence at multiple levels including token, type of event, visit information and temporal reference from anchor events.

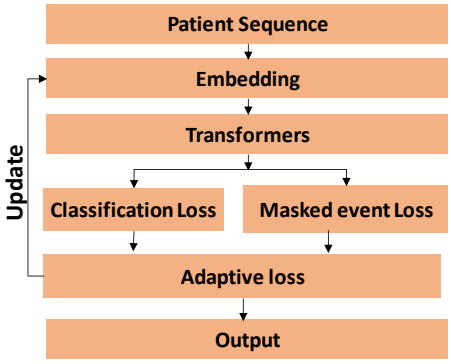

Figure 2: RareBERT architecture

Token embedding captures token level information or event information. *Type embedding* captures the event type information such as diagnosis, procedure or treatment. *Visit embeddings* capture patient visits or claims as patients could have many events within a visit. *Temporal embedding* captures time reference for an event from defined anchor date.

An example of RareBERT embedding encoding is presented in Figure 3 based on patient journey shown in Figure 1.

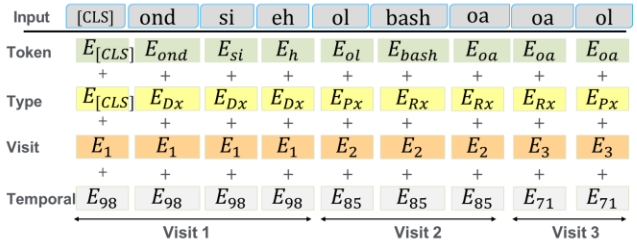

Figure 3: Illustration of input data for RareBERT

The RareBERT utilize only [CLS] token, while [SEP] token is ignored (Rasmy *et al.* 2020). RareBERT optimizes two loss function functions: (i) Masked event modelling; and (ii) event classification. The masked event modelling in sequence learning is an event fill-in-the blank task, where a model uses the event surrounding a mask token to predict the masked event which in turn helps in embedding generalization for different patient paths. The loss function weights for masked event modelling and event classification are updated adaptively to enhance the convergence.

Let $u = \{\boldsymbol{u_0}, \boldsymbol{u_1}, \boldsymbol{u_2}, \ldots \boldsymbol{u_{n-1}}\}$ represent input event tokens with n number of tokens. The RareBERT mask 15% of event token randomly with indices **m** and masked events are represented as $\boldsymbol{e}$ and non-masked event are represented as $\boldsymbol{e'}$. The mask event is replaced by a special token [MASK], random event from vocabulary, or unchanged with a probability of 80%, 10% and 10%, respectively. Let's $\boldsymbol{v} = \{v_0, v_1, v_2, \ldots v_{m-1}\}$ represent predicted events and $\boldsymbol{h}$ represents final hidden state of the first token [CLS] from RareBERT denote embedding for the whole sequence. The prediction for masked event modelling is optimized by minimizing negative log-likelihood (NLL) as shown below:

$$L_{MEM} = -E_{(u,v)\sim T} \log Pr(e'|\boldsymbol{h}) \ldots \quad (1)$$

where, $(\boldsymbol{u}, \boldsymbol{v})$ are pairs in the training dataset $\boldsymbol{T}$. The $P(e'|\boldsymbol{h})$ is computed as

$$P(e'|\boldsymbol{h}) = LogSoftmax(\boldsymbol{Q}\boldsymbol{h}) \quad \ldots (2)$$

where, $\boldsymbol{Q}$ is linear layer weight matrix. Similarly, the event classification module also utilizes the negative log-likelihood loss function as shown below

$$\boldsymbol{L_{EC} = -E_{(u,v)\sim T} \log Pr(c == 1|h)} \quad \ldots (3)$$

where, $\boldsymbol{c \in \{0, 1\}}$ with 1 representing positive class. The loss in RareBERT weighted combination of Eq. (1) and Eq. (3) and at i^{th} epoch is represented as

$$\boldsymbol{L_i = \alpha_i L_{EC,i} + L_{MEM,i}}$$

where, $\boldsymbol{\alpha_i}$ is weight at i^{th} epoch and is computed as

$$\boldsymbol{\alpha_i = min(\theta, L_{MEM,i-1}/(L_{EC,i-1} + \varepsilon))}$$

where, $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ are constant factor to control max bound of weights and correct for infinity scenarios. In this paper rare-patient identification set-up $\boldsymbol{\theta}$ is set based on class imbalance and $\boldsymbol{\varepsilon}$ is set to 10^{-6} close to handle any zero-division

error. Once RareBERT is trained embedding h is extracted and used within the PU-learning semi-supervised set-up for patient identification.

PU-Learning

RareBERT is used to learn the feature representation $\boldsymbol{h}$ which is used within the PU learning algorithm for patient identification**.** PU learning is a variation of the traditional set up where the training data consists of only positive and unlabeled examples where unlabeled examples include both positive and negative classes.

Let $\mathbf{P} \subset \mathbf{T}$ represent training set containing only positives patients and U$\subset$ **T** represents unlabeled classes where T is a universal set with all patients. The p and u are cardinality of P and U, respectively. The pseudo code for proposed PU learning based approach for patient identification is shown in Fig. 4.

Construct **P** and **U** by assigning $C_i = 0$ is h_i*is an unlabeled example*
While True:
  j=1
  Subset spies $\mathbf{S_j}$ where $\boldsymbol{S_j} \subset \boldsymbol{P}$ *and* $|\boldsymbol{S_j}| <$ | |
  Assign $C_i = 0$ where j $\subset \boldsymbol{S_j}$
  Train classifier to determine $Pr(C_i{=}1 \mid f(h))$ treating **U** as negative class.
  Evaluate $Pr(C_i{=}1 \mid f(h))$ for all observations
  Identify True Negative ($\mathbf{TN}_j$) where $\boldsymbol{TN_j} \subset \boldsymbol{U}$ where (C_i 1 ( )) $< \boldsymbol{\lambda}$
  Update **U** = **U** – $\boldsymbol{TN_j}$
  j = j+1

  # Stopping criteria
  if Recall > $\boldsymbol{\gamma}$ or $\boldsymbol{TN_j}$ - $\boldsymbol{TN_{j-1}} < \boldsymbol{\tau}$:
    break

$\boldsymbol{\lambda}$ represents *min* spies probability
LGBM (Ke *et al.* 2017*)* is used as classifier for training.

Figure 4: Pseudo code for PU learning

The PU learning is stopped based on two stopping criteria's: 1) recall threshold $\boldsymbol{\gamma}$ which represents the number of positives patients captured by classifier; and 2) True Negative (TN) threshold $\boldsymbol{\tau}$ which captures the minimum true negative identified during iteration.

Case Study

To illustrate the performance of RareBERT, X-linked hypophosphatemia (XLH) patients' identification is performed. XLH, is a condition that affects bones, muscles, and teeth due to the excessive loss of phosphate. Phosphate is lost through the urine, which causes low levels of phosphorus in the blood, a condition called phosphate wasting or hypophosphatemia. The XLH condition is selected due to: 1) high imbalance; 2) High proportion of un-diagnose patients with approximate prevalence of 1 in 20K patients in US; and 3) high disease burden (Skrinar *et al.* 2019).

Data Set-up

The analysis is performed using patient-level claims data with approximately 3.5 years of patient history from proprietary Symphony Health's IDV®[‡]. Patients with XLH and other comorbid conditions such as disorder of phosphorus metabolism, rickets, muscle weakness and, bone spurs were considered for further analysis.

In order to assure continuous patients' activity, standard eligibility criteria were applied. Eligibility criteria discard the patients who do not have a claim for a year. On top of it, a lookback period of two years was examined. Lookback was anchored on the first diagnosis of X-linked hypophosphatemia for XLH patients and on the last day of 2019 for patients with other comorbid conditions.

After applying comorbid condition, eligibility and lookback criteria, the final patient's cohort consists of 3,670 XLH patients and 263,187 unlabeled patients. The class imbalance between XLH and unlabeled patients is 1.37%.

Model Set-up and Feature Representation

For further modelling stratified sampling is performed to create a train (60%), validation (15%) and test (25%) dataset. Multiple approaches such as XGB with binary-based features, Long Short-Term Memory (LSTM), Med-BERT and RareBERT are developed for benchmarking. The first set of models are developed with classification set-up with unlabeled patient is treated as negative class. The hyper-parameter of XGB model is optimized using Bayesian optimization (Ruben, 2014). The LSTM model is set-up based on Kezi *et al.* (2019). The Kezi *et al.* 2019 takes LSTM embedding as input to GAN for further enrichment for patient identification. The Kezi *et al.* 2019 approach uses Word2Vec to encode event and pass to LSTM. The LSTM is set-up with 300 embedding size using word2vec, two layers and 256 hidden dimensions. Based on the initial set-up, AUPRC is compared across a set-up of models. The test performance for different models are reported in Table 1.

S. No.	Model	Test AUPRC
1	XGB (Binary-based feature)	43.0%
2	LSTM (Kezi *et al.* 2019)	43.3%
3	Med-BERT (Rasmy et al. 2020)	30.0%
4	RareBERT (w/ adaptive loss)	79.9%
5	**RareBERT (w/o adaptive loss)**	**80.1%**

Table 1: Performance on sequence dataset

[‡]Symphony Health, Integrated Dataverse (IDV)®, Sep. 1, 2019 – Dec. 31, 2019, unprojected de-identified patient Rx and medical claims, Jan 2020

Based on results from Table 1, XGB (with binary features) and LSTM are performing at an approximately 43% AUPRC. Med-BERT is performing worse, compared with the XGB (Binary-based feature) and LSTM for rare event scenarios. The RareBERT has shown an approximately 75% uplift in AUPRC performance, as compared with Med-BERT. The RareBERT shows similar performance with and without adaptive loss. The θ parameter for RareBERT (w/ adaptive loss) is set to 70 based on class imbalance observed in positive and unlabeled class.

Based on the above results, the RareBERT shows a strong feature representation capability to distinguish XLH and non-XLH patients. Additionally, the number of iteration and compute effort across models are compared and shown in Table 2. Adaptive loss has helped reduce compute effort by 66% as compared to without (w/o) an adaptive loss set-up of RareBERT.

S. No.	Model	# of Iterations	Time (hr)
1	XGB (Binary-based feature)	1700	3.25
2	LSTM (Kezi *et al.* 2019)	30	5.32
3	Med-BERT	56	22.32
4	RareBERT (w/o adaptive loss)	72	15.06
5	RareBERT (with adaptive loss)	29	5.15

Table 2: Med-BERT and RareBERT parameters

RareBERT and LSTM computation is performed on 2 GPU machine with 16 core and 2.5 GHz processor. The Med-BERT is computed on 4 GPU machine with 48 cores and 3.5 GHz processor. Although, Med-BERT is using higher configuration, it's more expensive than Rare-BERT with lower performance. The detailed set-up differences between Med-BERT and RareBERT is shown in Table 3.

Parameter	Med-BERT	RareBERT
Type of input code	ICD-9 + ICD-10 event CCS code	ICD-9 + ICD-10 event CCS code
Embeddings	Code + visit + Serialization	Code + code type + Visit + Temporal reference
Layer	6	2
Attention	6	6
Vector	32	16
Embedding size	192	96
classification loss	NLL	NLL
Masked event loss	NLL	NLL
Loss	Sum	Adaptive Weighted Sum
Event Prediction	Feed-forward Layer on averaged sequence	CLS Token

Table 3: Med-BERT and RareBERT parameters

Ablation study is performed on RareBERT architecture for key contribution of determine impact of individual component on final performance. The summary of ablation study is shown in Table 4. Check (✓) in Table 4 represent component that is considered for experimentation.

S. No.	Temporal	Type	CLS	Visit	Token	AUPRC	Δ from base
1*	✓	✓	✓	✓	✓	79.9%	
2		✓	✓	✓	✓	70.4%	9.5%
3	✓		✓	✓	✓	64.4%	14.9%
4		✓		✓	✓	56.0%	23.9%
5				✓	✓	42.3%	37.6%

** Base experiment*
***All experiments are performed with Adaptive loss*
****For experiment 4 & 5, CLS is replaced with feedforward layer*

Table 4: Ablation study summary of RareBERT

Based on Table 4, all Temporal, Type and CLS are significantly improving the performance of RareBERT architecture. Also, Adaptive loss helps to improve compute by ~66% based on Table 2. The next sub-section focuses on patient identification.

Patient Identification

The patients with a rare condition are identified using semi-supervised learning-based approach such as positive unlabeled (PU) learning. The PU learning framework as described in Fig. 4 is used with light gradient boosting machine (LGBM) as base classifier. The recall curves obtained from different models is shown in Fig. 5.

Based on Fig. 4, RareBERT shows a very high recall on test dataset with PU further boosting the recall performance with improved AUPRC to 79.8%. To further compare the performance of the model with semi-supervised learning where negative class is unknown pseudo F-1 score (Lee and Liu, 2003) is used. The pseudo F-1 score is defined as

$$Psuedo\ F1 - score = \frac{Recall^2}{\Pr(c == 1)}$$

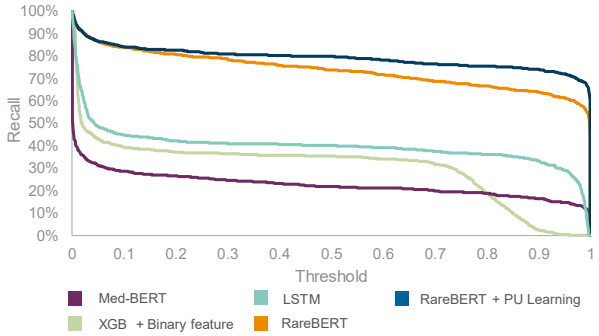

Figure 5: Recall across different models

The $Psuedo\ F1-score$ captures the performance of the current model with respect to the random baseline at defined threshold. Based on the above metric $Psuedo\ F1-score$ for all models are reported below:

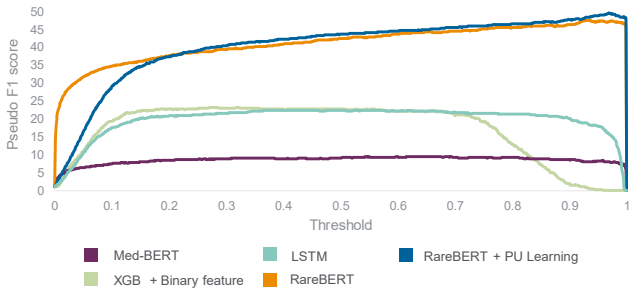

Figure 6: Psuedo-F1 score across different models

Based on Fig. 6, PU-BERT has shown an improvement of 4.4% in capturing unlabeled patients with rare conditions, compared with RareBERT. The max pseudo F1-score for a different model is reported in Table 5.

S. No.	Model	Max Pseudo F1
1	RareBERT	47.5
2	Med-BERT	9.5
3	Binary + XGB	23.2
4	LSTM	22.5
5	**PU + RareBERT**	**49.6**

Table 5: Max Pseudo F1-score to evaluate model lift

To understand the events driving the performance of RareBERT performance attention weights are extracted across different attention heads. Top 5 events leaning towards positive patient based on highest attention weights across different heads were the use of: 1) Vitamin D supplement; 2) Acidifying Agents (K-phos); 3) Opiate Agonists; 4) Anticonvulsants; 5) Beta-blocking agents.

Vitamin D and phosphate supplements are clinically recommended therapies for treating XLH and its symptoms (Dieter *et al.* 2019), thus correctly representing XLH patients. Similarly, Opiate Agonist is generally used as a pain killer. Skrinar *et al.* (2019) reported 97% of adults and 80% of children reported bone or joint pain/stiffness. Thus, the use of pain killer drugs might also be a proxy indicating pain related diagnosis for patients. Imel *et al.* (2012) report seizures as one of the symptoms for severe XLH patients which suggest the use of Anticonvulsants by XLH patients. The Beta-blocking agents are used in cardio conditions. There has been some research on the association of XLH and cardio vascular symptoms but not very concrete. However, model does suggest the usage of cardiovascular related drugs as one of the top indicators in identifying XLH and non-XLH patients.

Similarly, top events for negative patients were Chronic kidney disease, Deficiency and other anemia, Office visits for - interview, evaluation, consultation, Laboratory - Chemistry and Hematology procedures. Research suggests that patients with Chronic kidney disease and anemia also have low phosphorous levels, however, post dialysis most of the patients get their phosphorous levels back to normal (Nielsen *et al.* 2019) – this generally results in many cases where patients have got misdiagnosed as XLH or vice versa. However, RareBERT seems to have identified the differentiation between Chronic Kidney Disease and XLH patients. The top events associated with XLH and non-XLH patients, RareBERT seems to have picked the right signals thus boosting the confidence in the predictions of the model.

Model Sensitivity Analysis

One of the issues with claims dataset is noisy data coverage due to missing and wrong code capture (Tsang, 2020). Tsang (2020) summarized the drastic implications of noisy claims data on analysis. The current paper simulates the impact of missing claims through adversarial attack on the model. Simulation performed with 15% of events are randomly removed from patient journey and pre-trained model performance are re-evaluated on patient journey. The simulation is performed for 10 iteration; the average drop in AUPRC is reported in Figure 7.

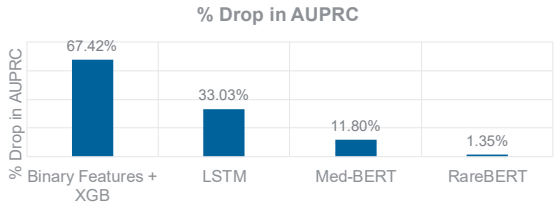

Figure 7: Percentage Drop in AUPRC

Based on Figure 7, RareBERT performance is quite stable even after 15% missing event noise induced in the dataset. Additionally, the impact on False Negative is more critical for patient identification, where we are tagging XLH patient as non-XLH. Results for False Negative (FN), *i.e.*, XLH patient is classified as non-XLH patient is shown in Figure 8.

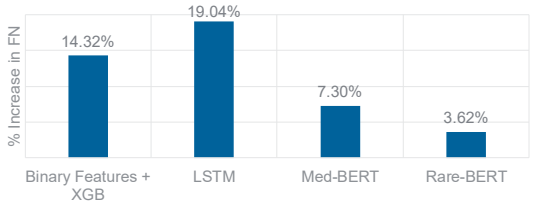

Figure 8: Percentage Increase in False Negative (FNs)

The above changes are evaluated with base model with precision set to 10%. The default output with 10% precision is reported below:

Model	Base AUPRC	TN	FP	FN	TP	Recall (%)
XGB (Binary-based feature)	43.0%	61040	4770	398	530	57.1
LSTM	43.3%	60725	5085	363	565	60.9
Med-BERT	30.0%	61931	3879	497	431	46.4
RareBERT *(with adaptive loss)*	**79.9%**	**58502**	**7308**	**116**	**812**	**87.5**

Table 6: Max Pseudo F1-score to evaluate model lift

The RareBERT has 87.5% recall with 10% precision to identify patients with rare-condition. Based on Table 5 RareBERT reports almost double potential patients with 87.5% recall as compared to Med-BERT.

To further evaluate consistency of the proposed RareBERT approach is also tested on Exocrine Pancreatic Insufficiency (EPI) rare condition. EPI is a condition in which the pancreas is not able to produce and/or transport enough digestive enzymes to break down food in the intestine.

There are two main challenges associated in identifying EPI patients within real world claims data are: 1) There is no specific ICD 9 diagnosis code for EPI, thus there is no direct way to identify patients who got diagnosed with EPI prior to October 2015 (ICD coding systems changed from ICD 9 to ICD 10 post October 2015); and 2) There are a good amount of patients who are misdiagnosed with a similar condition such as Abdominal pain, Chronic pancreatitis (CP), etc. These criteria are more aligned with the nuances associated with other rare disease conditions thus making it a non-trivial problem to solve.

The patients with relevant comorbid conditions are classified into EPI positive, negative and unlabeled are defined based on definition from Pyenson *et al.* 2019 as defined below:

- ***Positive class***: patients are classified into EPI positive if patient have filled minimum three prescription of PERT
- ***Negative class***: Patients are classified as EPI if fecal elastase-1 test is observed but PERT prescription is not filled
- ***Unlabeled class:*** patients with any of the relevant comorbid conditions.

After applying comorbid condition, eligibility and look-back criteria, the final patient's cohort consists of 15,045 EPI patients with 74,978 Negative patients. The dataset consists of 1.13 MM unlabeled patients with class imbalance between EPI positive and negative cases as 16.71%. The performance of models benchmarked against EPI dataset is shown in Table 7.

The results obtained are consistent with observed in XLH use case with RareBERT having 96.8% AUPRC. The XGB and LSTM based model shown a test AUPRC of 79% and 80.1%, respectively and MedBERT is unable to perform in highly imbalance scenarios.

S. No.	Model	Test AUPRC
1	XGB	79.0%
2	LSTM (Kezi *et. al.* 2019)	80.1%
3	Med-BERT (Rasmy *et. al.* 2020)	67.6%
4	Rare-BERT (w/ adaptive loss)	96.8%
5	Rare-BERT (w/o adaptive loss)	96.2%

Table 7: Performance on sequence dataset

Conclusion and Future Work

The paper presents a novel RareBERT architecture for feature representation in highly imbalanced longitudinal dataset. The learned feature representation is used with Positive Unlabeled (PU)-learning based approach for patient identification using patient administrative claims. The proposed PU learning algorithm utilizes positive class labels to identification potential patients with rare condition. The enhancement made in RareBERT helps it to learn representation better and converge faster based on adaptive loss.

Additionally, model sensitivity analysis is performed on the dataset to simulate partial data capture scenarios which is very prevalent across industries in transactional dataset. The simulations revealed that tree-based boosting is most susceptible to performance drop due to partial data capture. The performance drop within LSTM is also significant but significantly less as compared to tree-based bosting approach. The transformers-based model such as Med-BERT and RareBERT are using adversarial loss which helped them to generalize the feature representation better and have shown significantly more robust performance.

The RareBERT comes out to be most robust method. The robustness can be attributed to better generalization of temporal aspect for events and [CLS] token which is utilized for classification as compared to positional embedding and feed-forward layer used in Med-BERT.

The current experiment on RareBERT did not included any contextual information such as physical specialty or any patient related demographic information. The future experiments include bringing contextual information within the RareBERT architecture. This also includes developing a robust approach to handle differential Rx/Mx capture or integrating imputation mechanism to further strengthen the model performance in an unseen environment.

References

Bekker J. and Davis. J., 2018 Estimating the class prior in positive and unlabeled data through decision tree induction. *In AAAI*, pages 2712–2719

Colbaugh R., and Glass K. 2020. Finding Rare Disease Patients in EHR Databases via Lightly-Supervised Learning. *Technical Report, Volv Global, Lausanne, Switzerland*

Colbaugh, R *et al.* 2018. Learning to identify rare disease patients from electronic health records. *AMIA Annual Symposium*, San Francisco, CA USA, November 2018.

Dai D., and Hua S., 2016. Random under-sampling ensemble methods for highly imbalanced rare disease classification. *International Conference on Data Mining,* Barcelona, Spain, 12-15th Dec. 2016

Denis, F., Gilleron, R., Letouzey, F. 2005. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1), 70:83.

Du Plessis, M., Niu, G., Sugiyama, M. 2015. Convex formulation for learning from positive and unlabeled data. *International Conference on Machine Learning*, pp. 1386:1394.

Elliott E, Zurynski Y. 2015. Rare diseases are a 'common' problem for clinicians. *Aust Fam Physician.* 44:630–3

Elkan, C., Noto, K. 2008. Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213:220

Evans W, Rafi I. 2016. Rare diseases in general practice. *British J. General Practice*, Vol. 66.

Garg, R., S. Dong, S. Shah, and S. R. Jonnalagadda. 2016. A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records. *arXiv preprint arXiv*:1609.01586.

Hu Y., Chen F., Cai Y., and Yuan Y. 2019. A Random Under-Sampled Deep Architecture with Medical Event Embedding: Highly Imbalanced Rare Disease Classification with EHR Data. *International Conference on Data Science (ICDATA'19)*, July 29 - August 1, 2019. Las Vegas, NV

Ke G., *et al.* 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA

Pyeson B., *et al.* 2019. Applying Machine Learning Techniques to Identify Undiagnosed Patients with Exocrine Pancreatic Insufficiency. *Journal of Health Economics and Outcomes Research. Methodology and Health Care Policy,* 6(2).

Ruben M. C., 2014. BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits. *Journal of Machine Learning Research*, 15(Nov):3735--3739,

Tsang JP. 2020. Maladies of Claims Data: Manifestations, Origins, and Cures. *PMSA* Vol. 8(2).

Li W *et al.* 2018. Semi-supervised Rare Disease Detection Using Generative Adversarial Network. *Machine Learning for Health (ML4H) at NeurIPS*, 2018

Li Y., Rao *et al.* 2020. BEHRT: Transformer for Electronic Health Records. Scientific Reports, 7155

Madabushi *et al.* 2020, Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data. *arXiv preprint arXiv:2003.11563*

Rasmy L., Xiang Y., Xie Z., Tao C., and Zhi D. 2020. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *arXiv preprint arXiv: 2005.12833*

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L. 2017. Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*

Shang J, Ma T, Xiao C, Sun J. 2019. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:190600346.*

Skrinar A., *et al.* 2019. The Lifelong Impact of X-Linked Hypophosphatemia: Results From a Burden of Disease S urvey. *Journal of Endocrine Society,* 3(7): 1321–1334

Yu. K., *et al.* 2019. Rare Disease Detection by Sequence Modeling with Generative Adversarial Networks. *International Conference on Machine Learning,* Long Beach, California, PMLR, 2019.

Zurynski Y, et al. 2017. Rare disease: A national survey of pediatrician's experiences and needs. *BMJ Paediatrics,* Vol. 1

Lee, W.S., Liu, B. 2003. Learning with positive and unlabeled examples using weighted logistic regression. *ICML: Proceedings of the Twentieth International Conference on Machine Learning*

Haffner D., *et al.* 2019. Clinical practice recommendations for the diagnosis and management of X-linked hypophosphatemia. *Nature Reviews Nephrology,* 15, 435-455.

Skrinar, Alison, Dvorak-Ewell Melita, *et al.* 2019.The Lifelong Impact of X-Linked Hypophosphatemia: Results from a Burden of Disease Survey. *Journal of the Endocrine Society*, 3(7), 1321-1334.

Imel, Erik A and Econs, Michael J, 2012. Approach to the Hypophosphatemic Patient. *The Journal of Clinical Endocrinology and Metabolism,* 97(3):696.

Beck-Nielsen, *et al.* 2019. FGF23 and its role in X-linked hypophosphatemia-related morbidity. *Orphanet Journal of Rare Diseases*, 14(1):58.