Identifying Early Mpox Symptoms and Associations in Clinical Notes Using Natural Language Classification

Background/intro draft

The main objective of this project is to identify early unidentified or less associated symptoms of mpox infection among patients with related STD visits before the onset of main symptoms (e.g. mean incubation period 5.6 days, 95% credible interval 4.3-7.8) (1). In general, to gain insights about disease symptoms researchers have used a set of machine learning related approaches. Among them, random forest (RF) is a supervised machine learning method that can be used to predict a class or category based on a set of features (2). The application of these type of methods in conjunction with Natural Language Processing (NLP) that can capture disease symptoms from clinical notes remains largely unexplored. In the context of this project, we propose a natural language classifier that combines the NLP and RF to extract features and then use them for classification where the target output is to identify what symptoms contribute to the identification for mpox before the onset of the main symptoms.

Currently, several NLP model use the transformers architecture (e.g. Bidirectional Encoder Representation from Transformers [BERT]) because its state-of-the-art performance in many NLP tasks e.g. entity recognition, relation extraction, sentence similarity, natural language inference, and question answering. Transformers models are usually trained in two phases: the first part is to pretrain the model in large set of unlabeled data, the second is to fine-tune the model to the specific task with labeled trained data. A pretrained language can be fine-tuned to solve multiple task; this is known as transfer learning (3). Having a pretrained language in the public health toolbox ready to be deployed it is of the most important task for surveillance purposes. Specifically, at the onset of an emerging disease the model can be deployed for early detection. It can also support public health action through real time monitoring of anomalies in the data and population trends. Examples of transformers models applied in the health-related area include BioBERT(4), PubMedBERT(5), ClinicalBERT(6), and GatorTron (3).

Methods

Data set

The American Family Cohort (ABFM) is a collection of electronic health records that consist mainly of primary care practices in the United States. The cohort includes the visits of approximately 8 million patients and 12,000 clinicians. The collected data includes patients with sexually transmitted infectious (STI) visits between April 1, 2022, to April 1, 2023. The patient population consist of multiple races, ethnicities, and location.

Method overview

To identify less recognized symptoms of mpox from clinical notes, we define a matched cohort 1:2 (by encounter-month, gender, and age-group) of patients with an mpox diagnosis (ICD-10-DM B04) within the subset of patients with STI visits (Appendix 1). The matched groups are the mpox group which would be patients with a diagnosis code for mpox and the non-mpox group. The index date for the mpox group is the first mpox diagnosis found and for the non-mpox group is the matched STI visit to the mpox group.

Patient notes were collected between 1 and 16 days prior the index diagnosis for each patient in the cohort.

Once the notes are collected, they are preprocessed and broken down into sentences for analysis using a pre-trained transformer on the task of Name Entity Recognition (NER) able to identify symptoms and diagnosis from text. These featurization serves as input for a RF classifier and XGBoost algorithms. Using variable importance from RF we report the symptoms that support the classification results before the onset of the main symptom (rash) or formal mpox diagnosis.

Preprocessing (markup cleaning)

After cleaning the data markup (XLM, HTML, RTF) using regular expressions and the Spark NLP library (7), we use a sentence detection algorithm (7) adjust the patient notes to a length the pre-trained models can process. Three pretrained NER were used to identify symptoms and diagnosis from clinical notes: Bio-Epidemiology-NER (8), Biobert-ft-scispacy-NER and biobert_diseases_NER (9). For every patient, the identified entities were mapped as independent variables in the data set while maintaining record which patients were later diagnoses with mpox.

Classification

Using the mpox and non-mpox group in a single data set we implemented a classification problem where the outcome is mpox and the predictor variables are the symptoms and diagnosis identified by the NER. The data was split in 70% for training and 30% for testing. A Random Forest classifier was used for accuracy, robustness and flexibility when dealing with high dimensionality data, and its capability to report on feature importance. XGBoost was used for this classification step and was chosen because it is optimized for high performance computing and includes regularization parameters which simplifies the search space of this problem.

References

1. Madewell ZJ, Charniga K, Masters NB, Asher J, Fahrenwald L, Still W, et al. Serial Interval and Incubation Period Estimates of Monkeypox Virus Infection in 12 Jurisdictions, United States, May-August 2022. Emerg Infect Dis. 2023;29(4):818–21.
2. Biau G, Scornet E. A random forest guided tour. Test. 2016;25(2):197–227.
3. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. npj Digit Med. 2022;5(1):1–9.
4. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
5. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthc. 2022;3(1):1–24.
6. Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings ofthe 2nd Clinical Natural Language Processing Workshop. 2019. p. 72–8.
7. John Snow Labs. Spark NLP [Internet]. Spark NLP: State of the Art Natural Language Processing. 2022. Available from: https://sparknlp.org/
8. Raza S, Reji DJ, Shajan F, Bashir SR. Large-scale application of named entity recognition to biomedicine and epidemiology. PLOS Digit Heal [Internet]. 2022;1(12):e0000152. Available from:

http://dx.doi.org/10.1371/journal.pdig.0000152

9. Alonso Casero Á. Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature. E.T.S. de Ingenieros Informáticos (UPM); 2021.

Appendix 1. CCRS codes to define sexual transmitted infection visits

CCRS codes

CCRS category	**ICD-10**[1]	**description**
INF006	All related codes	HIV infection
INF007	All related codes	Hepatitis
INF010	All related codes	Sexually transmitted infections (excluding HIV and hepatitis)
FAC019		Socioeconomic/psychosocial factors
	Z206	Contact with and (suspected) exposure to HIV
	Z202	Contact with and (suspected) exposure to infections with a predominantly sexual mode of transmission
	Z113	Encounter for screening for infections with a predominantly sexual mode of transmission
	Z114	Encounter for screening for HIV
	Z1159	Encounter for screening for other viral diseases

[1] ICD-10-CM does not designate billing codes for PrEP, and the CCRS categories do not achieve this level of detail. The New York State Department of health and The New York City Department of Health and Mental Hygiene recommend the use of ZX related codes https://www.health.ny.gov/diseases/aids/general/prep/docs/icd_codes.pdf

Methods – patient sample characteristics

Patient sample: 549 patients, 183 with mpox diagnosis

- Clinical notes were collected between 1 and 15 days **prior** to the index diagnosis or the STI visit
- Label was assigned for those patients that had visits leading to an mpox diagnosis
- Not every mpox diagnosed patient had a note before the index diagnosis
- The number of identified patients with symptoms/disease prior index diagnosis, as a group entity, changes with the NER being used

	Mpox	Non-mpox	Total
Initial cohort	183	366	549
NER identified symptoms			
▪ Bio-Epidemiology-NER	133	314	447
▪ Biobertft-scispacy NER	65	338	403
▪ Biobert_diseases_NER	73	353	426

Results

- Random Forest variable importance was used to identify symptoms

NER name	Top 5 interpretable symptoms or associations	Accuracy	AUROC	Precision	Recall
Bio-Epidemiology NER	Communicable diseases, HIV, Hepatitis B, Hepatitis A, Arteriosclerosis	86%	0.92	83%	66%
Biobertft-scispacy NER	Constipation, myalgias, edema, abdominal pain, diarrhea	85%	0.77	0	0
Biobert_diseases_NER	Chest pain, abdominal pain, fatigue, fever,neumocystis carinii pneumonia	84%	0.88	9%	100%

Results

Classification results in **test set** for every combination of NER and random forest

Bio-Epidemiology-NER

		Predicted	
		Neg	Pos
Actual	Neg	85	5
	Pos	13	25

Biobert-ft-scispacy-NER

		Predicted	
		Neg	Pos
Actual	Neg	105	0
	Pos	19	0

Biobert_diseases_NER

		Predicted	
		Neg	Pos
Actual	Neg	107	0
	Pos	20	2

The Bio-Epidemiology-NER provides the best results so for far for this problem