A Novel Natural Language Processing and Machine Learning Classifier that Streamlines Extracting Drug-Adverse Event Data from Literature Reports

A Sorbello[1], R Hasan[2], H Francis[1], I Chang[2], M Ahadpour[1], M Laponsky[3], J Walsh[3], C Trier[3]

[1]CDER Office of Translational Sciences; [2]CDER Office of Computational Sciences; [3]Booz Allen Hamilton, Inc

Introduction

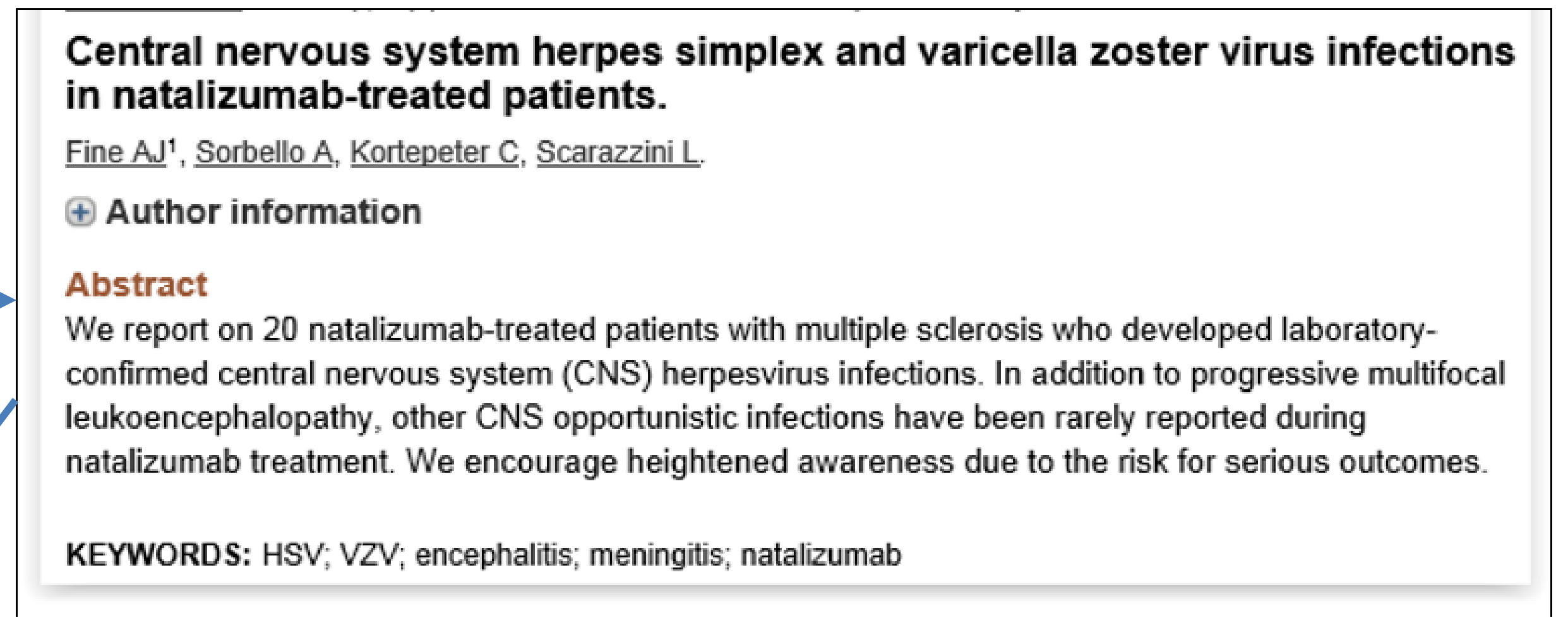
- Unintended harmful side effects of drugs contribute to increased ER visits, longer hospital stays, higher health care costs, and deaths.
- PubMed/MEDLINE is a leading scientific literature database produced by the National Library of Medicine that contains approximately 31 million citations
- CDER's Office of Translational Sciences developed a novel web-based literature mining application called PEARL to facilitate identifying adverse drug event (ADE) information from PubMed citations by leveraging the Medical Subject Heading (MeSH) indexing terms used to capture the clinical manifestations of these reactions[1]. PEARL is an acronym for 'Prospective Detection of Emerging Drug-Adverse Event Safety Signals from Relevant Scientific Literature through Quantitative Data Mining of MEDLINE Indexing Terms'.

Objectives and Challenges

- There is considerable drug safety data in free text format (title, abstract, key words) in unindexed PubMed/MEDLINE citations. We developed a novel natural language processing system to extract potential ADE information from these free text fields based on an annotated dataset of MEDLINE case reports[2], given that case reports are a frequent source of indexed ADE information in PubMed/MEDLINE.
- In leveraging MeSH indexing for pharmacovigilance, various features of its scope, granularity, and annotation rules limit our ability to extract data that may be useful for timely identification of newly emerging ADE safety signals from the most recently deposited PubMed/MEDLINE citations.

Methods

1. Our system employs a PubMed-trained language model using the Unified Medical Language System (UMLS) Metathesaurus. We constructed a text similarity index of all terms in the UMLS, enabling fuzzy detection of drug names and adverse event terms from selected free text fields within a PubMed citation. Candidate drug names were mapped to RxNorm, a standardized drug nomenclature, and candidate event terms were mapped to the Medical Dictionary for Regulatory Activities (MedDRA) terminology.



Central nervous system herpes simplex and varicella zoster virus infections in natalizumab-treated patients.

Fine AJ[1], Sorbello A, Kortepeter C, Scarazzini L.

⊕ Author information

Abstract

We report on 20 natalizumab-treated patients with multiple sclerosis who developed laboratory-confirmed central nervous system (CNS) herpesvirus infections. In addition to progressive multifocal leukoencephalopathy, other CNS opportunistic infections have been rarely reported during natalizumab treatment. We encourage heightened awareness due to the risk for serious outcomes.

KEYWORDS: HSV; VZV; encephalitis; meningitis; natalizumab

2. Each abstract is split into sentences, and co-occurring UMLS entity drug-event pairs are extracted as candidate causal pairs for model prediction.

We report on 20 natalizumab-treated patients with multiple sclerosis who developed laboratory-confirmed central nervous system (CNS) herpesvirus infections.

We report on 20 patients with multiple sclerosis who developed laboratory-natalizumab-treated confirmed central nervous system (CNS) herpesvirus infections.

In addition to progressive multifocal leukoencephalopathy, other CNS opportunistic infections have been rarely reported during natalizumab treatment.

3. Model predicts if a potentially causal linkage exists between a candidate drug-event pair.

- Natalizumab, herpesvirus infections

We trained an ADE classification model using PubMed/MEDLINE texts labeled with ADE pairs. The model is a fine-tuning task on the Biomed-RoBERTa model, which utilizes the Robustly Optimized BERT (RoBERTa) architecture and is pretrained on 2.68 million PubMed documents. We applied our base entity extraction system to a peer-reviewed dataset of 3000 ADE-annotated case reports, using annotated pairs of recognized drugs and adverse events as **positive** examples, and any additional drugs and adverse events detected within those texts as **negative** examples. Our resulting dataset contains 10442 drug-event co-occurrences, comprising 6431 ADE and 4011 non-ADE pairs across 4272 sentences. We trained our classifier on an 80% sample of this dataset to determine causal linkage between drug-event co-occurrences.

Results

The ADE classifier was tested on a dataset of 1,046 co-occurrences consisting of 656 ADE and 390 non-ADE pairs to evaluate model performance. This test dataset consisted of the remaining 20% of the originally generated dataset. The trained model achieved recall/precision/F1 scores of 0.91/0.915/0.91, respectively.

We assessed the feasibility of detecting ADE safety signals from free text titles, abstracts, and keywords in PubMed/MEDLINE citations in the use case of the sodium glucose co-transporter 2 (SGLT2) inhibitors. In total, 2,291 citations pertaining to these drugs were assessed, spanning from February 2008 to August 2020.

For the four FDA-approved SGLT2 inhibitors, the table below displays selected candidate ADE pairs corresponding with known side effects as described in their respective product package inserts.

Selected drug-adverse event pairs that correspond to known labeled toxicities for four drugs chosen for evaluation

Drug Name	Drug_CUI	Adverse Event	Event_CUI
canagliflozin	C2974540	Hypoglycemia	C0020615
dapagliflozin	C2353951	Hypoglycemia	C0020615
empagliflozin	C3490348	Hypoglycemia	C0020615
ertugliflozin	C4079805	Hypoglycemia	C0020615
canagliflozin	C2974540	Genital infection	C0729552
dapagliflozin	C2353951	Genital infection	C0729552
ertugliflozin	C4079805	Genital infection	C0729552
canagliflozin	C2974540	UTI	C0042029
dapagliflozin	C2353951	UTI	C0042029
canagliflozin	C2974540	Ketoacidosis diabetic	C0011880
dapagliflozin	C2353951	Ketoacidosis diabetic	C0011880
empagliflozin	C3490348	Ketoacidosis diabetic	C0011880

CUI = Concept unique identifier; UTI = urinary tract infection

Conclusions and Limitations

CONCLUSIONS

- This novel ADE classifier provides a feasible framework to extract ADE information from free text fields in PubMed/MEDLINE citations.
- ADE safety signals illustrative of known labeled toxicities can be detected. The classifier system may potentially detect novel emerging ADEs.

LIMITATIONS

- False positives caused by drug-indication and other non-drug safety mentions
- Complexities in cross-terminology mapping due to the highly granular UMLS and MedDRA terms.

FUTURE ENHANCEMENTS

- Methodological refinements to increase precision in identifying valid candidate ADE pairs, especially filtering out drug-indication pairs.
- Compare PEARL outputs that leverage MeSH indexing for signal detection to the new classifier to assess their strengths and limitations.
- Assess the classifier for detecting new emerging safety issues with a larger group of drug products

References

1. Sorbello A, Ripple A, Tonning J, et. al. Harnessing scientific literature reports for pharmacovigilance. Prototype software analytical tool development and usability testing. Appl Clin Inform 2017 Mar 22;8(1):291-305. PubMed PMID: 28326432.

2. Gurulingappa H, Rajput A, Roberts A, et. al. Development of a benchmark corpus to support automatic extraction of drug-related adverse effects from medical case reports, J Biomed Inform 2012 Oct; 45(5):885-892. PubMed PMID: 22554702.

The views expressed are those of the authors and do not represent those of the US FDA or the US Government.