Natural Language Processing Methods for Detecting Non - Therapeutic Drug Use in Clinical Notes

Nikki Adams, PhD
Rihem Badwe, PharmD
Division of Health Care Statistics
October 22, 2024

The findings and conclusions in this presentation are those of the authors and do not necessarily present the official position of the Centers for Disease Control and Prevention.

Project Background and Goals

Current Project

- Develop an algorithm using hospital care data, both medical codes and clinical notes, to detect non-therapeutic stimulant and opioid use:
  - Prescription misuse
  - Use of illicit drugs
  - Non-therapeutic use of some unspecified drug
- Support was provided for the development of the Stimulant Algorithm by the Department of Health and Human Services' Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) under the project titled "Utilizing Natural Language Processing and Machine Learning to Enhance the Identification of Stimulant and Opioid-Involved Health Outcomes in the National Hospital Care Survey."

Past Projects

- **NCHS also received funding from OS-PCORTF for fiscal years 2018 and 2019**
  - FY2018 – Opioid use and opioid overdose
    - Enhancing Identification of Opioid-involved Health Outcomes Using National Hospital Care Survey Data (White et. al. 2021)
  - FY2019 – Selected mental health issues and substance use disorders
    - Identifying Co-occurring Disorders Among Patients With an Opioid-involved Hospital Encounter Using National Hospital Care Survey Data (Brown et. al. 2022)

Data Source

The National Hospital Care Survey

- The National Center for Health Statistics (NCHS) conducts the National Hospital Care Survey (NHCS).
  - NHCS is a nationally-representative sample of 608 non-institutional, non-federal hospitals with at least six staffed inpatient beds .
  - Collects one year of emergency department (ED) visits and inpatient hospital discharges.

The National Hospital Care Survey

- Collects Uniform Billing (UB)-04 administrative claims data or electronic health records (EHR) data from participating hospitals
- Collects patient personally identifiable information (PII) data
- Collects medical codes and labels for conditions, labs, and procedures (all sources) and medications (EHR only)
- Notes (only for EHR data) contain information typically thought of as “clinical notes” as well as text labels from medical codes.

https://www.cdc.gov/nchs/nhcs/index.html

Annotation and Creating the Gold Standard Dataset

Creating a standard for development and testing

- ‘Ground truth’ dataset, containing truth about concepts of interest
- Determinations made by clinical experts (the annotators)
- Can be used for development, training, or testing
- The algorithm has two components, the results of which are integrated to produce a final algorithm
  - Code component to analyze medical codes
  - Natural language processing (NLP) component to analyze clinical notes.
- Annotators examined data used in both components (medical codes and clinical notes)

Overview of the making of the gold standard dataset

- Consider variables to be included in the final dataset
- Construct form that allows relevant information to be recorded
- Sample from the survey, targeting concepts of interest
- Have domain experts (clinicians) trial the form and questions, and come to consensus during training
- Once training shows consensus, multiple annotators work independently

Clinical Notes Structure and Format

Types of EHR data for clinical notes

- These two types of EHR have different characteristics
  - EHR data that NCHS collects directly from hospitals
  - EHR data acquired from the American College of Emergency Physicians (ACEP)

Two types of EHR data for notes

One note record

One encounter (ID 123X)

ENCOUNTER ID	SOURCE	SETTING	NOTE TYPE	NOTE NUMBER	NOTE TEXT
456Y	Direct EHR	Inpatient	Medication	1	<?xml version=”1.0”>text xmlns
456Y	Direct EHR	Inpatient	Social History	2	<?xml version=”1.0”>text xmlns
123X	ACEP	ED		1	Patient has medical history of asthma hypertension…LabsWBC.5.norma
123X	ACEP	ED		2	Albuterol& Flonaise Advair…Alcohol use Cannabis Speed Crack…

Structure and format affect approach

- **Direct EHR:**
  - Depending on one's project needs, XML might have to be parsed. This was not necessary for this project
  - Metadata from the Note Type gives a lot of information
  - Non-therapeutic status of a drug could be inferred with high accuracy

Structure and format affect approach

- **Data from ACEP:**
  - Without metadata, that simple formula from direct EHR is not possible
  - Considered trying to deduce the note type by either rule-based NLP or with a model, but discovered that a single note record could have multiple types of information in it
  - Needed a different approach: determine the non-therapeutic status by a machine learning model

A Machine Learning Approach

The first model: logistic regression

- **We used a logistic regression machine learning model from the scikit learn package in Python by:**
  - Creating a snippet of text surrounding the drug term
  - Annotating a few hundred snippets of text for non-therapeutic status (yes/no)
  - Performing simple cleaning and normalization of the data

Cleaned snippet	Label
…with a history of DRUGTERM use disorder …	1
… as needed. DRUGTERM 28 mg film .. (MOTRIN) 400 mg	0

Results

- **Training and testing on a 75/25 training/test split**

Precision[1]	Recall[2]	F1[3]
0.86	0.96	0.91

[1] Also known as positive predictive value

[2] Also known as sensitivity

[3] Harmonic mean of precision and recall

Incorporating the model into algorithm

- We used rules to narrow the space as much as possible
- We used regular expression to discount lab reports, drug screenings, and general templates for risk assessments.
- Then:

Patient PII Disclosure Issues with First Model

- The classic way of encoding words in that model involves a dictionary of all words seen during training
- The clinical notes are PII data, meaning names could be in the dictionary
- This cannot be released to the public

The second model: A fine-tuned BERT model

- We used the same snippets from training from the first model
- Results were not great in the first pass
- Annotated more to get ~800 examples, including approx. 200 altered or fabricated examples
- Trained and tested on an 80/20 split

Precision	Recall	F1
0.90	0.93	0.92

Next Steps

Dissemination

- Dr. Rihem Badwe presents in more detail on annotation work here at FCSM on October 24.
- Research file in NCHS Research Data Center for results of algorithm applied to National Hospital Care Survey 2020
- Methodology report with detail on methods expected 2025 (will appear here):
  - https://www.cdc.gov/nchs/products/series/series02.htm
- GitHub releases for algorithm planned:
  - https://github.com/CDCgov

Questions

Thank you for your attention and feedback!

Presenter Contact Information

- Nikki Adams – Nikki.Adams@cdc.hhs.gov
- Rihem Badwe – Rihem.Badwe@cdc.hhs.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov