Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Harnessing natural language processing and machine learning to enhance identification of opioid-involved health outcomes in the national hospital care survey

by Amy M. Brown and Nikki Adams

Release date: October 22, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Harnessing Natural Language Processing and Machine Learning to Enhance Identification of Opioid-involved Health Outcomes in the National Hospital Care Survey

Amy M. Brown, MPH, Nikki Adams, PhD[1]

Abstract

To build data capacity and address the U.S. opioid public health emergency, the National Center for Health Statistics received funding for two projects. The projects involve development of algorithms that use all available structured and unstructured data submitted for the 2016 National Hospital Care Survey (NHCS) to enhance identification of opioid-involvement and the presence of co-occurring disorders (coexistence of a substance use disorder and a mental health issue). A description of the algorithm development process is provided, and lessons learned from integrating data science methods like natural language processing to produce official statistics are presented. Efforts to make the algorithms and analytic datafiles accessible to researchers are also discussed.

Key Words: Opioids; Co-Occurring Disorders; Data Science; Natural Language Processing; Hospital Care

1. Introduction

1.1 Project Background

The Centers for Disease Control and Prevention (CDC)'s National Center for Health Statistics (NCHS) is one of 13 federal statistical agencies and serves as the principal health statistics agency in the United States. Data collection systems include vital records, population surveys, and provider surveys. The National Hospital Care Survey (NHCS) is designed to provide healthcare utilization data on hospital-based settings. During fiscal years 2018 and 2019, NCHS received funding from the U.S. Department of Health and Human Services Office of the Secretary Patient Centered Outcomes Research Trust Fund (OS-PCORTF) for two projects to build data capacity and integrate data science methods to help address the opioid public health emergency by improving the enumeration and characterization of opioid-involved hospital encounters. The projects involved developing two sets of algorithms that use all available structured and unstructured data elements in the increasingly large and complex NHCS data. The first algorithm is designed to identify hospital encounters that involve past or present use of opioids and the specific type of opioid taken. The second algorithm is designed to identify opioid users with co-occurring disorders, specifically a co-existing substance use disorder (SUD) and mental heath issue (MHI). These enhanced algorithms are designed to identify cases that may have been missed by earlier algorithms that relied solely on medical codes to identify behavioral health issues of interest in submitted survey data.

1.2 Data Sources

The current NHCS sample includes 608 non-institutional, non-federal hospitals with six or more staffed inpatient beds. Participating hospitals are asked to submit data on all hospitalizations and emergency department visits within each calendar year. Hospitals can choose between submitting data electronically in one of two main formats: (1) Uniform Bill (UB)-04 administrative claims or (2) electronic health records (EHRs). Submitted data covers a wide

[1]Amy M. Brown, MPH, National Center for Health Statistics, 3311 Toledo Road, USA, 20782 (wri1@cdc.gov); Nikki Adams, PhD, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, USA, 20782 (oxf7@cdc.gov). All views expressed in this article are those of the authors and do not necessarily reflect those of the Centers for Disease Control and Prevention or National Center for Health Statistics.

range of data elements, including structured data like standard medical codes (e.g., diagnosis and procedure codes) and unstructured data representing providers' clinical notes. The data for both PCORTF projects is from the 2016 NHCS, which includes approximately 9.6 million inpatient and emergency department encounters from 158 hospitals.

2. Methods

2.1 Algorithm Development Process

The algorithm development process began with creating case definitions for each of the two algorithms. The Enhanced Opioid Identification Algorithm case definition included criteria to identify past or present use of all prescription and illicit opioids and substances with opioid-like effects, such as kratom. The case definition for the Co-occurring Disorders algorithm has criteria to identify past or present use disorders for all substances and selected mental health issues related to anxiety, depression, obsessive compulsive disorder, suicidality and trauma and stressor related disorders.

Both algorithms used two components that make use of all available data submitted for each encounter to identify cases. The first component searched for relevant medical codes and keywords across all fields with medical codes and associated labels or descriptions. In the second component, a sample of clinical notes were annotated, or labeled, to inform development of natural language processing (NLP) and machine learning techniques to analyze unstructured clinical text notes. The NLP component involved breaking up clinical note text into sentences and searching sentence-by-sentence to detect several types of rule-outs, such as excluding opioid, SUD or MHI mentions associated with dates later than 2016. Negation detection was also performed to determine whether a mention of drug use, finding or disease mentioned within a narrative medical report was being asserted or denied (e.g., "no history of anxiety"). Misspellings of opioid mentions were also found by using a machine learning model named entity recognition (NER), that was trained to recognize drug names and then compare detected drug names to the spelling of opioid names.

Detailed methodology for the Enhanced Opioid-Identification Algorithm is described in a published report (White et. al., 2021). A similar report describing detailed methodology for the Co-occurring Disorders Algorithm is forthcoming. We are currently conducting a study to validate performance of these enhanced algorithms. Findings from this validation study, in progress, will be used to make modifications and produce a final set of algorithms.

2.2 Medical Code Component

The 2016 survey year marked the first time that hospitals participating in the NHCS had the option to submit EHR data, which presented the opportunity to search for additional types of medical codes that were not available in previous years. The medical code component of both algorithms searched for selected diagnosis, medication, procedure, and laboratory test codes from standard code systems. In some instances, selected keywords were also searched for in the code label or description if the hospital used a non-standard code system. The final medical code and search term lists can be accessed on the NCHS Research Data Center website (NCHS, 2020; NCHS, 2021).

2.3 NLP Component

The introduction of EHR data in the 2016 NHCS also made unstructured data from clinical notes available for the first time in the survey's history. These clinical notes may provide greater specificity regarding the type of opioid taken compared to the standardized medical codes assigned to encounter. Similarly, searches of clinical note sections like "Past Medical History" and "Social History" can help identify patients who were clinically diagnosed with a SUD or MHI prior to the encounter.

The broad methodology of the NLP component was primarily rule-based, with a machine learning component implemented to detect misspellings of opioid medications and implemented in Python. First, there was an initial period of gathering keywords and phrases targeting categories of interest, and these were refined over time. Text was extracted from encounters that passed an initial note-level filtering (for example, filtering out generic patient information pamphlet-style notes). The text was then broken into sentences and sentence-level exclusions were also

performed (for example, filtering out sentences about family history). Then keyword searches were performed. For opioid detection only, named entity recognition (NER) was used to identify drug names not already on our list (misspellings), followed by a separate process of automatically comparing candidate terms' spelling to known opioids and other drugs. If similarly spelled, these were set aside to be confirmed by human annotators that they indeed were a misspelling of an opioid. The NER was trained in-house using an annotated dataset from the 2016 NHCS on top of a base English model from the spaCy software package (spaCy, https://spacy.io). All keyword matches were then passed through a negation detection filter, Negex (Chapman et. al. 2001), to determine if they should be excluded due to negation (for example, "denies abusing opioid medication"). Non-negated matches were mapped to umbrella variable categories to produce the final dataset. For more information, see White et. al. 2021.

The addition of these clinical notes to the 2016 NHCS, however, also posed new challenges. Only 8.7% of all encounters had at least one clinical note record available for use in the NLP component. Clinical notes were optional to submit as they are more challenging for hospitals to extract. Submitted clinical notes also had misspellings, truncation, white space variations and varied in formatting (free text or XML) requiring extensive data cleaning and reformatting. Lastly, the NLP component had to be developed in a closed computing environment that complied with applicable data security rules but limited available software and computing power. The closed environment also became difficult to access during the COVID-19 pandemic because it could not be accessed from home during telework. To accommodate the closed environment and limited software availability, project staff relied more heavily on internally developed software, such as developing our own annotation software solution.

3. Results

3.1 Annotation Results

The annotation dataset was partitioned into development sets to build the NLP component and evaluation sets to measure its performance, one development and evaluation set each for the opioid-involvement algorithm, the MHI algorithm, and the SUD algorithm. Results of the annotation reflect the performance of the medical code component, NLP component and the full algorithm with both components combined to distinguish between the presence or absence of opioids, SUDs, and MHIs as identified by the annotators. All annotated encounters had clinical notes, so both components of the algorithm could be applied equally to all encounters in the annotation dataset. Tables 3.1-1 to 3.1-3 present the following performance metrics:

- Recall: Percentage of correctly identified positives out of all true positives, also known as sensitivity
- Precision: Percentage of identified positives that are true positives, also known as positive predictive value (PPV)
- F1: Harmonic mean of recall and precision, a common measure of algorithm performance
- MCC: Matthew's Correlation Coefficient, also known as Pearson's Phi Coefficient, provides a measure balanced over true and false negatives and positives

For the Enhanced Opioid Identification Algorithm, recall, F1 and MCC scores were low for the medical code component, while the full algorithm (which combined both the medical code and NLP components) exhibited the best overall performance with an F1 of 92.5% and MCC of 0.77 (Table 3.1-1). This result was anticipated because the clinical notes often mention opioid drugs, particularly therapeutics, for which there is no corresponding medical code in any of the coded fields. In the Co-occurring Disorders Algorithm, both components combined performed the best in identifying MHIs with an F1 of 95.2% and MCC of 0.82 (Table 3.1-2). For SUDs, the medical code component had the highest performance (F1 of 94.9% and MCC of 0.80), but the combined algorithm still performed well with a relatively high F1 (90.2%) and MCC (0.80) (Table 3.1-3).

Table 3.1-1
Performance of the Enhanced Opioid Identification Algorithm compared to the annotated data set by component

	Medical Code Component	NLP Component	Full Algorithm (Code & NLP Components)
Recall	25.5%	94.8%	96.9%
Precision	96.9%	88.5%	88.6%
F1	20.4%	91.5%	92.5%
MCC	0.30	0.74	0.77

NOTES: NLP, natural language processing; F1, harmonic mean of recall/precision; MCC, Matthew's Correlation Coefficient.
SOURCE: National Center for Health Statistics, National Hospital Care Survey, 2016

Table 3.1-2
Performance of the Co-Occurring Disorders Algorithm to identify MHI compared to the annotated data set by component

	Medical Code Component	NLP Component	Full Algorithm (Code & NLP Components)
Recall	86.7%	74.7%	93.3%
Precision	99.2%	96.6%	97.2%
F1	92.5%	84.2%	95.2%
MCC	0.77	0.58	0.82

NOTES: MHI, mental health issue; NLP, natural language processing; F1, harmonic mean of recall/precision; MCC, Matthew's Correlation Coefficient.
SOURCE: National Center for Health Statistics, National Hospital Care Survey, 2016

Table 3.1-3
Performance of the Co-Occurring Disorders Algorithm to identify SUD compared to the annotated data set by component

	Medical Code Component	NLP Component	Full Algorithm (Code & NLP Components)
Recall	91.6%	90.2%	99.3%
Precision	98.5%	81.1%	82.6%
F1	94.9%	85.4%	90.2%
MCC	0.90	0.70	0.80

NOTES: SUD, substance abuse disorder; NLP, natural language processing; F1, harmonic mean of recall/precision; MCC, Matthew's Correlation Coefficient.
SOURCE: National Center for Health Statistics, National Hospital Care Survey, 2016

Codes were highly precise but tended to have lower recall. In particular, as the case definition for opioid-involvement was any opioid use, keyword searches for opioids worked well here, while there was frequently no accompanying billing code for that use. False negatives and false positives for the NLP component arose for a variety of reasons. A fuller understanding of those reasons is still developing as we continue error analysis and continue analyzing the

results of a follow-on validation study. However, initial investigation reveals reasons for false negatives such as drug interactions being characterized as "abuse" by diagnosis code but "use" in the notes, where drug use, by our case definition, did not rise to the level of a substance use disorder. A frequent source of false positives in the NLP portion was for nicotine, where the case definition for nicotine SUD changed between the time of annotation and the time of algorithm completion. Being a former smoker was initially excluded, but later included. That, combined with a "former smoker" status being mentioned in the notes but not accompanied by a diagnosis code of personal history of nicotine dependence, resulted in the NLP component flagging former smokers more often than the medical code component. Analysis of errors and the results of our validation study is ongoing.

3.2 Application of Algorithm to the 2016 NHCS

Following annotation, the enhanced algorithms were applied to the 2016 NHCS dataset. Tables 3.2-1 and 3.2-2 present results for all encounters (irrespective of whether they included clinical notes) as alongside results for only encounters with at least one clinical note record. The Enhanced Opioid Identification Algorithm identified a total of 1,370,827 opioid-involved encounters. One fifth (20.3%) of encounters were identified exclusively by the NLP component and would have been missed by an algorithm relying solely on medical codes. When restricted to encounters with available clinical notes, a small percentage of encounters (0.9%) were only identified by the medical code component (Table 3.2-1).

The Co-occurring Disorders Algorithm identified a total of 659,225 opioid-involved encounters as SUD-only, MHI-only or co-occurring disorders. Precisely 10.3% of encounters were identified solely by the NLP component. When restricted to encounters with available clinical notes, 10,542 encounters (6.0%) were only identified by the medical code component (Table 3.2-2).

Table 3.2-1
Number and percentage of opioid-involved encounters with and without available clinical notes identified by medical code-based, NLP and both algorithms in the National Hospital Care Survey 2016 dataset

	Over all encounters		Over only encounters with clinical notes	
	Counts	Percentages	Counts	Percentages
Code component alone	1,060,495	77.4%	2,819	0.9%
NLP component alone	277,958	20.3%	277,958	88.8%
Both code and NLP component	32,374	2.4%	32,374	10.3%
Total	1,370,827	100.0%	313,151	100.0%

NOTES: NLP, natural language processing.
SOURCE: National Center for Health Statistics, National Hospital Care Survey, 2016

Table 3.2-2
Number and percentage of SUD-only, MHI-only or co-occurring encounters with and without available clinical notes identified by medical code-based, NLP and both algorithms in the NHCS 2016 dataset

	Over all encounters		Over only encounters with clinical notes	
	Counts	Percentages	Counts	Percentages
Code component alone	494,458	75.0%	10,542	6.0%
NLP component alone	67,584	10.3%	67,584	38.6%
Both code and NLP component	97,183	14.7%	97,183	55.4%
Total	659,225	100.0%	175,309	100.0%

NOTES: NLP, natural language processing.
SOURCE: National Center for Health Statistics, National Hospital Care Survey, 2016

4. Discussion

4.1 Lessons Learned on Building Data Science Capacity

NLP can be a valuable tool in analyzing submitted hospital data even when standardized medical codes are available, as these codes may come with many logistical challenges. However, proper handling of varied data formats is not trivial and software management may rely on factors outside of a researcher's control, as with operating in a closed environment. Additionally, NLP may increase recall, but this is often at the cost of lower precision.

Algorithm performance metrics for both projects emphasize the importance of using a comprehensive approach to identifying health concepts of interest in administrative claims and EHR data such as opioid-involved and co-occurring disorders. We found algorithms that use both medical code searches and text mining techniques performed best overall and identified encounters that otherwise would have been missed.

Data collection systems used to produce official statistics are handling increasingly complex and large volumes of data. The challenges faced in this study with incorporating these methods demonstrate the importance of building a data analytics platform that has sufficient computing power for big data, is easily adaptable for installation of the best available software solutions and enables all designated users to access it from approved devices both onsite and remotely. In addition, providing respondents with tools to extract and submit data in a standardized format may also reduce burden and improve data quality. NCHS has developed an HL7 Implementation Guide that can be implemented by a hospital or EHR vendor that identifies desired data elements and preferred formats (HL7 International, 2021).

4.2 Next Steps for the Project

Analytic datasets with enhanced algorithm variables have been made available to the research community in the NCHS Research Data Center. More information on the datasets can be found here: https://www.cdc.gov/rdc/b1datatype/dt1224h.htm. The Python code for the NLP component will also be available in a forthcoming posting to https://github.com/oxf7 by Spring 2022.

A validation study involving abstraction of selected encounters by external annotators has been conducted and we are currently analyzing algorithm errors based on those results. The algorithm will be refined and an updated version released.

References

Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 34(5):301–10. DOI: 10.1006/ jbin.2001.1029

HL7 International. HL7 CDA® R2 Implementation Guide: National Health Care Surveys (NHCS), R1 STU Release 3 – US Realm.

National Center for Health Statistics Research Data Center. (2020). Linked Data on Hospitalizations, Mortality, and Drugs: Data from the National Hospital Care Survey 2016, National Death Index 2016-2017, and the Drug Involved Mortality 2016-2017. Available from: https://www.cdc.gov/nchs/data/nhcs/Task-3-Doc-508.pdf.

National Center for Health Statistics Research Data Center. (2021). Identifying Co-Occurring Disorders among Opioid Users Using Linked Hospital Care and Mortality Data: Capstone to an Existing FY18 PCORTF Project. Available from: https://www.cdc.gov/nchs/data/nhcs/FY19-RDC-2021-06-01-508.pdf.

SpaCy: Industrial-strength natural language processing (NLP) with Python and Cython. Available from: https:// github.com/explosion/spacy-models/releases/tag/ en_core_web_md-2.2.5

White, D.G., Adams, N.B., Brown, A.M., O'Jiaku-Okorie, A., Badwe, R., Shaikh, S., Adegboye, A. (2021b). Enhancing identification of opioid-involved health outcomes using National Hospital Care Survey data. National Center for Health Statistics. Vital Health Stat 2(188). https://www.cdc.gov/nchs/data/series/sr_02/sr2-188.pdf.