1	Title: Development and Evaluation of Machine Learning Models for the Detection of Emergency
2	Department Patients with Opioid Misuse from Clinical Notes
3	
4	Authors: Usman Shahid ^{1,2} , Natalie Parde ^{1,2} , Dale L. Smith ^{1,3} , Grayson Dickinson ⁴ , Joseph
5	Bianco ⁵ , Dillon Thorpe ^{1,3} , Madhav Hota ⁶ , Majid Afshar ⁷ , Niranjan S. Karnik ^{1,3,8} , Neeraj
6	Chhabra ^{1,8,9*}
7	
8	1. AI.Health4All Center for Health Equity using Machine Learning and Artificial Intelligence,
9	College of Medicine, University of Illinois Chicago, Chicago, IL, USA
10	2. Natural Language Processing Laboratory, Department of Computer Science, University
11	of Illinois Chicago, Chicago, IL USA
12	3. Department of Psychiatry, University of Illinois Chicago, Chicago, IL USA
13	4. William Carey University, College of Osteopathic Medicine, Hattiesburg, MS,
14	USA
15	5. University of Illinois College of Medicine, Chicago, IL USA
16	6. University of Missouri-Kansas City, Kansas City, MO, USA
17	7. Department of Medicine, School of Medicine and Public Health, University of Wisconsin
18	Madison, Madison, WI, USA
19	8. Institute for Research on Addictions, University of Illinois Chicago, Chicago, IL USA
20	9. Department of Emergency Medicine, University of Illinois Chicago, Chicago, IL
21	USA
22	
23	Corresponding Author: Neeraj Chhabra, MD, MSCR. Department of Emergency
24	Medicine, University of Illinois Chicago. 808 S Wood Street, 4 th floor, Chicago, IL 60612
25	Neeraj1@uic.edu

\sim	\sim
• •	6
	C)
_	~

27	Grants: This work was supported by the following grants from the National Institute on Drug
28	Abuse (NIDA)/NIH: K23DA055061 (Chhabra), R01DA051464 (Afshar), R61DA057629
29	(Karnik/Chhabra) and, in part, by the National Center for Advancing Translational Sciences
30	(NCATS), through Grant UL1TR002003. The content is solely the responsibility of the authors
31	and does not necessarily represent the official views of the National Institutes of Health.
32	
33	Abstract
34	
35	Objectives
36	The accurate identification of Emergency Department (ED) encounters involving opioid misuse
37	is critical for health services, research, and surveillance. We sought to develop natural language
38	processing (NLP)-based models for the detection of ED encounters involving opioid misuse.
39	
40	Methods
41	A sample of ED encounters enriched for opioid misuse was manually annotated and clinical
42	notes extracted. We evaluated classic machine learning (ML) methods, fine-tuning of publicly
43	available pretrained language models, and a previously developed convolutional neural network
44	opioid classifier for use on hospitalized patients (SMART-AI). Performance was compared to
45	ICD-10-CM codes. Both raw text and text transformed to the United Medical Language System
46	were evaluated. Face validity was evaluated by term feature importance.
47	
48	Results
49	There were 1123 encounters used for training, validation, and testing. Of the classic ML
50	methods, XGBoost had the highest AU_PRC (0.936), accuracy (0.887), and F1 score (0.863)
51	which outperformed ICD-10-CM codes [accuracy 0.870; F1 0.830]. Logistic regression, support

It is made available under a CC-BY-NC-ND 4.0 International license .

- 52 vector machine, and XGBoost models had higher AU_PRC using transformed text, while
- 53 decision trees performed better using raw text. Excluding XGBoost, fine-tuned pre-trained
- 54 language models outperformed classic ML methods. The best performing model was the fine-
- 55 tuned SMART-AI based model with domain adaptation [AU_PRC 0.948; accuracy 0.882; F1
- 56 0.851]. Explainability analyses showed the most predictive terms were 'heroin', 'opioids',
- 57 'alcoholic intoxication, chronic', 'cocaine', 'opiates', and 'suboxone'.
- 58
- 59 Conclusions
- 60 NLP-based models outperform entry of ICD-10-CM diagnosis codes for the detection of ED
- 61 encounters with opioid misuse. Fine tuning with domain adaptation for pre-trained language
- 62 models resulted in improved performance.

63

It is made available under a CC-BY-NC-ND 4.0 International license .

64 Introduction

65

66 Drug overdose is the leading cause of accidental death in the United States with the majority 67 involving an opioid.¹ A critical healthcare setting for the initiation of treatments and medications for opioid use disorder (OUD) is the Emergency Department (ED).² While treatments exist to 68 69 decrease mortality related to opioid use, only a minority of patients with OUD are engaged in medical treatment.^{3,4} As people with OUD disproportionately utilize emergency services, ED 70 71 encounters serve as valuable opportunities for interventions and linkage to outpatient healthcare 72 resources. The accurate identification of patients at risk for OUD in the ED setting is critical to 73 providing these life-saving treatments. 74 75 Current methods to identify patients with opioid misuse, defined as taking an opioid in a manner 76 other than prescribed or using illicit opioids, rely on clinical interactions and documentation, 77 often in the form of diagnosis codes.^{5–8} Universal manual screening for opioid misuse in the ED has been proposed but is highly resource-intensive and infrequently performed.^{5,9} Prior work 78 79 has shown that documentation of opioid misuse by International Classification of Diseases, 10th revision, Clinical Modification (ICD-10-CM-CM) diagnosis code is highly insensitive and fails in 80 81 identifying a large proportion of patients that would benefit from interventions.¹⁰ The low 82 sensitivity of ICD-10-CM based approaches for patient identification has impacts on research, 83 surveillance, resource allocation, and clinical services in the healthcare setting. Methods in 84 natural language processing (NLP) and machine learning (ML) to process clinical notes for 85 clinical workflows have shown promise to build tools for screening opioid misuse in the inpatient setting.^{11–13} Such models rely on inpatient hospital documentation from the electronic health 86 87 record (EHR), which far exceeds the amount of documentation performed during a typical ED 88 encounter. It is currently unknown if NLP models trained using machine learning can be 89 successfully domain adapted for screening opioid misuse in the ED setting.

It is made available under a CC-BY-NC-ND 4.0 International license .

90

91 The goal of the current study is to develop and evaluate NLP-based machine learning models to 92 screen patients for opioid misuse during their ED encounter. We hypothesized that these 93 models would outperform diagnosis codes in the identification of patients with opioid misuse, 94 potentially highlighting the utility of such models for health services delivery, research, and 95 surveillance. 96 97 Methods 98 99 Setting and Cohort Development 100 University of Illinois Hospital and Health Sciences System (UIHealth) is comprised of a 462-bed 101 tertiary referral hospital with a 30-bed emergency department providing care across 45,000 102 patient encounters annually. It is located on the westside of Chicago, Illinois within an urban 103 area that has among the highest opioid-related mortality in the United States.^{14,15} UIHealth has 104 maintained Epic as its EHR vender since 2020 with a clinical data warehouse (CDW) of all 105 patient encounters and associated data. The retrospective observational cohort of ED 106 encounters used for training and testing of classifier models was drawn from the notes and 107 reports extracted from the CDW with inclusion criteria of age greater than or equal to 18 years 108 and encounter origination in the ED. A train/validation/test corpus was developed using a 109 sample of approximately 1200 ED encounters. The sample was enriched for suspected opioid 110 misuse using previously described methods involving sampling ED encounters with a positive 111 urine opiate screen without coexisting opioid prescription or medication administration on 112 hospital intake or an opioid-related ICD-10-CM diagnostic code and matching with encounters 113 lacking these criteria.^{10,16} Enrichment was pursued to allow a balanced dataset for more efficient 114 model training and improved classification. Encounters matching the criteria during the study 115 period of September 2020 to March 2023 were identified and matched in a 1:1 ratio on

It is made available under a CC-BY-NC-ND 4.0 International license .

disposition status (hospital admission or discharge) with encounters lacking previously
described criteria or additional criteria placing the patient at risk for opioid misuse: an ICD-10CM code for a chronic pain diagnosis, an order for naloxone, or an order for a urine drug
regardless of result. We matched on disposition in lieu of age and sex to minimize bias from
increased documentation, testing, and entry of ICD-10-CM diagnosis codes based on admission
status and because age and sex are potentially important predictor variables for opioid misuse.

122

123 Manual Annotation

124 The dataset was further processed with expert manual annotations. As opioid misuse 125 represents a heterogeneous pattern of drug use and is difficult to categorize solely from 126 variables within the EHR, potential cases and non-cases were manually annotated in a blinded 127 format using a structured manual annotation schema. For the purposes of annotation, opioid 128 misuse was defined as taking an opioid for reasons other than prescribed or as an illicit drug. 129 consistent with definitions from the National Institute of Drug Abuse (NIDA) and the National 130 Survey on Drug Use and Health (NSDUH).⁶⁻⁸ Annotators underwent a one-hour online training 131 session with an expert in emergency addictions care (NC). They were required to achieve an 132 interrater reliability kappa of greater than 0.80 with the expert prior to independent annotation. 133 Annotators again completed cases in parallel with the expert and were given additional cases to 134 annotate only after achieving threshold interrater reliability following every 100 cases 135 independently annotated. Case details were extracted from the EHR by the annotators into a structured REDCap data collection form for each encounter.^{17,18} 136 137

Using previously described methods, the presence of opioid misuse was determined using a 5point Likert scale indicating the probability of opioid misuse which included the categories of definite, highly probable, probable, definitely not, and uncertain.^{10,11} Probable cases required either: 1) history of opioid misuse evident in clinical notes but no current documentation for the

It is made available under a CC-BY-NC-ND 4.0 International license .

encounter; 2) provider mention of drug-seeking behavior; or 3) evidence of other drug misuse,
except alcohol, in addition to prescription opioid use. Highly probable cases had either more
than one probable case criteria or provider mention of suspicion of opioid misuse. Definite cases
were classified by either patient self-report of opioid misuse or documentation by provider of
patient misusing an opioid. Cases lacking any of the above criteria were classified as 'definitely
not.' For classification, cases annotated as probable, highly probable, or definite were
categorized as exhibiting opioid misuse.

149

150 Model Development

151 Encounters comprising the cohort were sorted chronologically by encounter date and divided 152 into training, validation, and testing sets at a ratio of 70/15/15. We evaluated multiple model 153 architectures, all of which can be broadly categorized as either classic (feature-based) machine 154 learning and neural methods. Classic machine learning methods included logistic regression, 155 support vector machines, decision trees, and eXtreme Gradient Boosted (XGBoost) trees. 156 Neural methods utilized publicly available pre-trained language models or a previously trained convolutional neural network opioid classifier for use on hospitalized patients (SMART-AI).¹³ We 157 158 experimented with two variants of the SMART-AI classifier. Firstly, we used the classifier as a 159 feature extractor and trained an MLP classifier on these features. In the second variant, we fine-tuned this model end-to-end on our dataset to improve domain adaptation.¹⁹ The pre-160 161 trained language models included Bidirectional Encoder Representations from Transformers 162 (BERT), BioBERT, Longformer, Clinical Longformer, and Clinical BigBird. BERT is an encoder-163 only transformer model developed by Google and Longformer is a modified transformer model 164 which can account for longer text sequences for use on general text.^{20,21} BioBERT. Clinical 165 Longformer, and Clinical BigBird are domain-adapted language models pre-trained on medical corpora.^{22,23}. For neural models, we performed domain adaptation with fine-tuning of the pre-166 167 trained model on the training data by using the averaged embeddings from the final hidden layer

It is made available under a CC-BY-NC-ND 4.0 International license .

of the pre-trained model and added a Multi-Layer Perceptron classifier on top with one hidden
layer of size 256 and dropout layers with probability 0.3. Rectified Linear Unit (ReLU) was used
as the activation function between intermediate layers and final probabilities were obtained
using the sigmoid function, similar to the binary output provided by a logistic regression
classifier.

173

174 Text Processing

175 All clinical documents originating from individual ED encounters were concatenated in 176 chronological order, including notes from staff with direct interaction with patients in the ED such 177 as physicians and nurses. For each machine learning method, we evaluated both the use of raw 178 text as well as transformed text (feature engineering), where applicable. For classic machine 179 learning methods, text went through minimal preprocessing by converting to lower case and 180 breaking the text into unigram (single word, top 10000 unigrams by frequency) feature vectors 181 to represent individual variables. No feature engineering was performed for fine-tuning of pre-182 trained language models and the natural language of the text was used up to the context length 183 of the language model (e.g., 512 for BERT based and 4096 for Longformer based models)

184

185 In addition to the raw text features, the text was also mapped to medical concepts and 186 converted into structured codes, referred to as concept unique identifiers (CUIs). The text was 187 mapped to the National Library of Medicine's Unified Medical Language System (UMLS) 188 concept unique identifier (CUI) codes using the clinical Text And Knowledge Extraction System 189 (cTAKES). cTAKES is an open-source NLP engine which identifies named entities in raw text 190 and maps them to the UMLS.^{24,25} This transformation accounts for negation and outputs CUI 191 codes. An additional feature of cTAKES transformation is the stripping of protected health 192 information from the raw text as part of processing into named entities and CUIs as well as 193 bringing together semantically similar terms to the same medical concept using the UMLS

It is made available under a CC-BY-NC-ND 4.0 International license .

194 Metathesaurus (i.e, opioid use disorder and OUD). We chose to evaluate both raw and

transformed text. This is because raw text retains the contextual information in a sentence,

196 while the transformed text using cTAKES maps similar terms to a single CUI thus negating

some variability due to individual clinician vocabulary and word choice.

198

199 Statistical Analysis

200 Models were compared along multiple metrics with the primary being Area Under the Precision 201 Recall Curve (AU PRC). The AU PRC considers both precision (positive predictive value) and 202 recall (sensitivity) and was chosen as the primary metric given the importance of identifying all 203 positive cases and better accounting for imbalanced datasets. We also considered other metrics 204 such as accuracy, F1 score, area under the receiver operator curve (AU ROC), and language 205 model complexity as determined by the number of parameters. We evaluated the number of 206 parameters (i.e., model weights) for language models since a higher number of parameters can 207 represent a need for more significant computational resources. We selected our final model as 208 the one with the most favorable metrics, prioritizing AU PRC. In the case of similarly performing 209 models, parsimony was prioritized. We also compared model performance against clinical detection of opioid misuse ICD-10-CM diagnosis codes.^{26,27} Entry of ICD-10-CM codes 210 211 represent clinician identification of opioid misuse. ICD-10-CM codes are commonly used for 212 problem list generation, billing, surveillance, and research. They represent a baseline that any 213 potential classification model should outperform. As a binary variable, we only determined a 214 subset of metrics for ICD-10-CM code performance: precision, recall, F1 score, and accuracy. 215

216 Feature/Variable Importance

To evaluate face validity of the final model, we estimated feature importance for the bestperforming model on the held-out test set. Given the black box nature of many neural methods,
we used Local Interpretable Model-agnostic Explanations (LIME) to evaluate the top 25 features

It is made available under a CC-BY-NC-ND 4.0 International license .

which were most predictive of opioid misuse.²⁸ Designed for explainable artificial intelligence,
the LIME algorithm fits multiple surrogate linear regression models to approximate a machine
learning model and evaluates feature importance for each prediction in the form of beta
coefficients, referred to as LIME scores. We used LIME to evaluate the features which predicted
whether a patient was positive or negative for opioid misuse for each observation in the held-out
test set.

226

All analyses were performed in Python (version 3.12) using PyTorch (version 2.4).²⁹ This study was reviewed and exempted from review as non-human subjects research by the institutional review board of the primary institution. The study conforms, where appropriate, to Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence (TRIPOD+AI) guidelines (Appendix A).³⁰

232

233 Results

234 Of the initial 1200 annotated ED encounters comprising the study cohort, 77 encounters were

removed for enhanced privacy protections associated with patient chart (62) or lack of any text

documentation (15), leaving 1123 encounters for training, validation, and testing. Of the total

237 1123 cases, 570 were labeled as opioid misuse positive and 553 labeled negative.

238 Demographic and other information for the cohort is shown in Table 1. Encounters were divided

temporally into training, validation, and testing sets with 786, 168, and 169 encounters,

respectively. In the test set, there were 75 encounters labeled positive for opioid misuse and 94labeled negative.

242

243 Table 1. Demographics of Cohort

Variable	n (%)

Age	
18-34	285 (25.38%)
35-49	255 (22.71%)
50-64	360 (32.06%)
65+	223 (19.86%)
Sex	
Female	543 (48.35%)
Male	580 (51.65%)
Race/Ethnicity	
Non-Hispanic Black	636 (56.63%)
Hispanic or Latino	294 (26.18%)
Non-Hispanic White	159 (14.16%)
Other	25 (2.23%)
Patient Declined	9 (0.80%)
Insurance	
Medicaid	665 (59.22%)
Medicare	245 (21.82%)
Private	130 (11.58%)
Other	18 (1.60%)
No Information	65 (5.79%)
ED Disposition	
AMA	77 (6.9%)
Admit	487 (43.4%)
Discharge	417 (37.13%)
Expired	2 (0.18%)

It is made available under a CC-BY-NC-ND 4.0 International license .

Observation	59 (5.25%)
Sent to operating room	2 (0.18%)
Other	79 (7.0%)

Notes: ED, Emergency Department; LWBS, left without being seen by medical provider; AMA,
left against medical advice, eloped, or left without being seen by provider. "Other" disposition
category includes transfers to other facilities, transfers to behavioral health, transfers to
obstetrics, and unspecified dispositions.

248

249 Of the classic machine learning methods, XGBoost had the highest AU-PRC (0.936), accuracy

250 (0.8876), and F1 score (0.8633) on the held-out test set (Table 2). Logistic regression, support

vector machine, and XGBoost-based models all had higher AU_PRC when used with CUIs,

while only decision tree models had improved metrics with untransformed text. Of the classic

253 machine learning methods, only XGBoost models outperformed ICD-10-CM codes in recall, F1

score, and accuracy.

255

256 With the exception of XGBoost, the neural classifiers utilizing pre-trained language models

257 generally outperformed the classic machine learning methods in AU_PRC. The best performing

258 models were the SMART-AI based models which used CUIs as inputs. The SMART-AI model

that underwent transfer learning with domain adaptation through fine tuning with frozen CUI

260 embeddings had the highest AU_PRC and AU_ROC of all models evaluated.

261

262 Table 2: results of classifiers by method and data preparation

263

Method	Data	Precision*	Recall*	F1*	Accuracy*	AU_ROC	AU_PRC
ICD Codes	ICD	0.9643	0.7297	0.8308	0.8698	-	-

	Codes						
Logistic	CUIs	0.831	0.7867	0.8082	0.8343	0.8831	0.9036
Regression							
Logistic	Text	0.7763	0.7867	0.7815	0.8047	0.8955	0.8786
Regression							
Support	CUIs	0.8551	0.7867	0.8194	0.8462	0.9055	0.908
Vector							
Machine							
Support	Text	0.6988	0.7733	0.7342	0.7515	0.8383	0.8088
Vector							
Machine							
Decision	CUIs	0.8133	0.8133	0.8133	0.8343	0.8322	0.8548
Tree							
Decision	Text	0.8101	0.8533	0.8312	0.8462	0.8469	0.8643
Tree							
XGBoost	CUIs	0.9375	0.8	0.8633	0.8876	0.9268	0.936
XGBoost	Text	0.9016	0.7333	0.8088	0.8462	0.9356	0.9336
BERT	Text	0.8333	0.6667	0.7407	0.7929	0.9094	0.9101
BioBERT	Text	0.7465	0.7067	0.726	0.7633	0.8465	0.8182
Longformer	Text	0.9667	0.7733	0.8593	0.8876	0.8738	0.9087
Clinical	Text	0.9231	0.8	0.8571	0.8817	0.8894	0.9164
Longformer							
Clinical	Text	0.9104	0.8133	0.8592	0.8817	0.918	0.9329
Bigbird							
SMART-AI	CUIs	1	0.64	0.7805	0.8402	0.8753	0.8934
SMART-AI	CUIs	0.9828	0.76	0.8571	0.8876	0.939	0.9419
as Feature							

It is made available under a CC-BY-NC-ND 4.0 International license .

Extractor**							
SMART-AI	CUIs	0.9661	0.76	0.8507	0.8817	0.9464	0.9476
Finetuned***							

264 * Threshold for classification is P(Opioid) ≥ 0.5 for all classifiers except SMART-AI base variant

265 (for which use their recommended 0.05)

266 ** Only the linear layers were trained, leaving the rest of the model frozen

- 267 *** All parameters were fine-tuned apart from frozen CUI embeddings
- 268
- 269 We considered model complexity as well, and we present those findings in Table 3. The
- 270 SMART-AI based models had fewer parameters than all pretrained language models.
- 271 Explainability analyses using LIME show the CUIs more predictive of opioid misuse were
- 272 'heroin', 'opioids', 'alcoholic intoxication, chronic', 'cocaine', 'opiates', and 'suboxone' (Figure 1).
- 273
- 274 Table 3: Neural Classifier Parameters

Model	Data	No. of Parameters
BERT	Text	109 M
BioBERT	Text	108 M
Longformer	Text	148 M
Clinical Longformer	Text	148 M
Clinical Bigbird	Text	127 M
SMART-AI	CUI	12.15 M
SMART-AI as Feature Extractor	CUI	12.4 M
SMART-AI Finetuned	CUI	12.4 M

275

276 Figure 1: Feature Importance for predicting opioid misuse

It is made available under a CC-BY-NC-ND 4.0 International license .



277

*Descriptions along the y-axis are of the following format: concept unique identifier (text
descriptions from the National Library of Medicine's Unified Medical Language System, number
of encounters from the training set with the concept unique identifier, proportion of encounters
with the concept unique identifier that was labeled as displaying opioid misuse). LIME scores
range from 0-1 with higher scores indicating higher predictive power on the test set.

283

284

285 Discussion

286 In this evaluation of model-based detection of patient encounters with opioid misuse in the ED 287 setting, the best performing NLP classifiers outperformed ICD-10-CM codes for the detection of 288 opioid misuse. The best performance was noted from the models which were trained for the 289 detection of opioid misuse in hospitalized patients and adapted to this domain by employing 290 transfer learning techniques. These models were also more lightweight, based on the number of 291 parameters, than those using pretrained language models. In general, models based on text 292 transformed to a standard UMLS lexicon performed as well or better than those using raw text in 293 comparable settings. Some of the pretrained language models, especially Clinical Bigbird, had

It is made available under a CC-BY-NC-ND 4.0 International license .

comparable performance, most likely due to its large context window and pretraining on medical
text (MIMIC-III clinical notes).²³ It is important to note that XGBoost, a lightweight classifier,
stood out among classical machine learning methods by delivering comparable performance to
neural approaches without necessarily requiring text pre-processing with cTAKES.

298

299 Model-based detection of patients with opioid misuse has implications in healthcare delivery. 300 disease surveillance, and research. The accurate detection of patients likely to benefit from 301 opioid-specific interventions is critical in the ED setting. Universal manual screening is 302 infrequently performed due to being resource intensive and because it may not be ideally suited 303 for the emergency setting where there are multiple competing clinical concerns in a highly time-304 sensitive environment.⁹ Diagnosis code entry, which serves as a surrogate marker for the detection of opioid misuse, is highly insensitive.^{10,31,32} Current methods of patient detection 305 306 would leave many patients that would likely benefit from medical treatment for opioid use 307 disorder without being detected. As medical treatment is already uncommon among patients 308 with opioid misuse in the ED, improved detection is critical.³³ Similarly, surveillance for opioid 309 misuse among ED patients has implications for resource allocation and programmatic funding. 310 Underdetection leaves already resource-strapped environments with a diminished ability to 311 advocate for expanded resources necessary to care for patients with opioid use disorder. As 312 ICD-10-CM codes are highly relied upon for cohort development in research, the improved 313 detection with NLP-based models would allow for decreased selection bias and improved 314 validity of research reliant on diagnosis codes for cohort discovery.

315

316 In the current study, we evaluated both the use of raw text and transformed text as features for 317 candidate models. Both have theoretical benefits and drawbacks. Raw text, when used with pre-318 trained transformer-based models, has the benefit of context from surrounding words to improve 319 classification. It also avoids an additional layer of computational burden that occurs with text

It is made available under a CC-BY-NC-ND 4.0 International license .

320 pre-processing. In our analysis, we evaluated transformed text to a standardized lexicon from 321 UMLS using cTAKES. Although context is lost by using cTAKES as it transforms named entities 322 without regard for nearby terms, advantages of using cTAKES are that it has the ability to: 1) 323 map disparate terminology to single distinct concepts; and 2) remove identifying protected 324 health information (PHI) such as names and dates. The benefits of mapping to CUIs negate 325 some of the effects of individual variations in vocabulary, allowing similar concepts to be 326 aggregated. The removal of PHI carries implications for model deployment as multiple 327 institutions can theoretically share lists of deidentified CUIs from clinical encounters for model 328 processing, allowing for the accurate surveillance for disease processes by public health entities 329 across multiple institutions without compromising patient privacy. 330

331 Among similarly performing models, the simpler model is preferred due to explainability, 332 scalability, and transportability. Pretrained language models, while computationally expensive, 333 did not outperform the SMART-AI classifier, which has fewer parameters by an order of 334 magnitude. In addition, the baseline SMART-AI classifier without adaptations was outperformed 335 by a version fine-tuned on ED-specific data, which speaks to the successful adaptation to this 336 domain using transfer learning from external sources. Note that most of the SMART-AI 337 parameters (approximately 11.2M) were pre-trained CUI embeddings. Evaluation of black-box 338 neural network-based clinical models for interpretability is an important consideration prior to 339 clinical deployment. The domain-adapted SMART-AI model underwent evaluation for 340 interpretability with the most important CUIs in predicting opioid misuse representing concepts 341 with high face validity. Some terms with decreased association with opioid misuse, such as 342 'Positron Emission Tomography,' are likely related to features noted in sparse training data for 343 either the original or adapted classifier or a result of term overlap where some strings may be 344 mapped to incorrect concepts. The next steps before clinical use of a domain-adapted SMART-345 Al model involve subgroup validation and recalibration on a larger cohort of ED encounters.

2	Λ	ົ
ັ	4	υ

347	All experimental results should be interpreted within the context of their limitations. As ICD-10-
348	CM codes were a portion of the criteria for cohort development, performance metrics for ICD-
349	10-CM codes were likely inflated. For experiments with pretrained language models, we were
350	limited by the context window limitations of the individual models. Preprocessing of text with
351	cTAKES requires a license and carries the potential limitation of mapping of named entities to
352	incorrect CUIs. All models designed for clinical use should undergo bias and fairness
353	assessments to ensure equity. Bias assessments for the current study were unrevealing owing
354	to the small test set size and should be evaluated on a larger sample with appropriate mitigation
355	techniques prior to clinical deployment.
356	
357	Conclusion
358	Natural language processing-based models can outperform entry of ICD-10-CM diagnosis
359	codes in the detection of encounters with opioid misuse in the ED setting. Fine tuning with
360	domain adaptation for pre-trained language models resulted in improved performance of our
361	opioid misuse classifier.
362	
363	

It is made available under a CC-BY-NC-ND 4.0 International license .

364 References

365

- 366 1. Abuse NI on D. Drug Overdose Death Rates | National Institute on Drug Abuse (NIDA)
- 367 [Internet]. 2024 [cited 2024 Aug 5]; Available from: https://nida.nih.gov/research-
- 368 topics/trends-statistics/overdose-death-rates
- 2. D'Onofrio G, McCormack RP, Hawk K. Emergency Departments A 24/7/365 Option for
- 370 Combating the Opioid Crisis. N Engl J Med 2018;379(26):2487–90.
- 371 3. Dowell D, Brown S, Gyawali S, et al. Treatment for Opioid Use Disorder: Population
- 372 Estimates United States, 2022. Morb Mortal Wkly Rep 2024;73(25):567–74.
- Shastry S, Manini AF, Richardson LD, Lin MP. US ED Opioid-Related Visits Increase, While
 Use of Medication for Opioid Use Disorder Undetectable, 2011-2016. J Gen Intern Med
 2020;35(3):965–6.
- 376 5. Chalmers CE, Mullinax S, Brennan J, Vilke GM, Oliveto AH, Wilson MP. Screening Tools
- 377 Validated in the Outpatient Pain Management Setting Poorly Predict Opioid Misuse in the
- 378 Emergency Department: A Pilot Study. J Emerg Med 2019;56(6):601–10.
- 379 6. Abuse NI on D. Summary of Misuse of Prescription Drugs [Internet]. Natl. Inst. Drug Abuse.
- 380 -- [cited 2023 Jan 30];Available from: https://nida.nih.gov/publications/research-
- 381 reports/misuse-prescription-drugs/overview
- 382 7. Commonly Used Terms | Opioids | CDC [Internet]. 2023 [cited 2023 Oct 5]; Available from:
 383 https://www.cdc.gov/opioids/basics/terms.html
- 384 8. 2015 National Survey on Drug Use and Health: Methodological Summary and Definitions |
 385 CBHSQ Data [Internet]. [cited 2023 Jan 30];Available from:

It is made available under a CC-BY-NC-ND 4.0 International license .

386 https://www.samhsa.gov/data/report/2015-national-survey-drug-use-and-health-

- 387 methodological-summary-and-definitions
- 388 9. Chhabra N. Death by a Thousand Screens: A Practical Role for Machine Learning in
 389 Emergency Medicine. Ann Emerg Med 2023;82(4):531–2.
- Chhabra N, Smith D, Pachwicewicz P, et al. Performance of International Classification of
 Disease-10 codes in detecting emergency department patients with opioid misuse. Addict
 Abingdon Engl 2023;
- 393 11. Sharma B, Dligach D, Swope K, et al. Publicly available machine learning models for
- identifying opioid misuse from the clinical notes of hospitalized patients. BMC Med Inform

395 Decis Mak [Internet] 2020 [cited 2021 Jan 17];20. Available from:

- 396 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7191715/
- 397 12. Afshar M, Sharma B, Bhalla S, et al. External validation of an opioid misuse machine
- learning classifier in hospitalized adult patients. Addict Sci Clin Pract 2021;16(1):19.
- 399 13. Afshar M, Sharma B, Dligach D, et al. Development and multimodal validation of a
- 400 substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI):
- 401 a retrospective deep learning study. Lancet Digit Health 2022;4(6):e426–35.
- 402 14. Anthony K. Cook County closes 2020 with record highs of 875 gun-related homicides, 1,599
- 403 opioid deaths [Internet]. Chic. Sun-Times. 2021 [cited 2021 Jan 26];Available from:
- 404 https://chicago.suntimes.com/metro-state/2021/1/2/22210281/cook-county-homicide-total-
- 405 2020-970-opioid-covid
- 406 15. •. As Opioid Overdose Deaths Hit New Record, Pressure Grows for Safe Places to Inject
 407 Drugs in Chicago [Internet]. NBC Chic. 2022 [cited 2024 Oct 23];Available from:

It is made available under a CC-BY-NC-ND 4.0 International license .

408 https://www.nbcchicago.com/news/local/as-opioid-overdose-deaths-hit-new-record-

- 409 pressure-grows-for-safe-places-to-inject-drugs-in-chicago/2730602/
- 410 16. Afshar M, Joyce C, Dligach D, et al. Subtypes in patients with opioid misuse: A prognostic
- 411 enrichment strategy using electronic health record data in hospitalized patients. PloS One

412 2019;14(7):e0219717.

- 413 17. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data
- 414 capture (REDCap)--a metadata-driven methodology and workflow process for providing
- 415 translational research informatics support. J Biomed Inform 2009;42(2):377–81.
- 416 18. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international
 417 community of software platform partners. J Biomed Inform 2019;95:103208.
- 418 19. Dligach D, Afshar M, Miller T. Pre-training phenotyping classifiers. J Biomed Inform
 419 2021;113:103626.
- 420 20. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional
- 421 Transformers for Language Understanding [Internet]. 2019 [cited 2023 Jun 8];Available
 422 from: http://arxiv.org/abs/1810.04805
- 423 21. Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer [Internet].
- 424 2020 [cited 2024 Oct 30];Available from: http://arxiv.org/abs/2004.05150
- 425 22. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation
- 426 model for biomedical text mining [Internet]. 2019 [cited 2024 Oct 30];Available from:
- 427 http://arxiv.org/abs/1901.08746

- 428 23. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-Longformer and Clinical-BigBird:
- 429 Transformers for long clinical sequences [Internet]. 2022 [cited 2024 Oct 30];Available from:
- 430 http://arxiv.org/abs/2201.11838
- 431 24. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge
- 432 Extraction System (cTAKES): architecture, component evaluation and applications. J Am
- 433 Med Inform Assoc JAMIA 2010;17(5):507–13.
- 434 25. Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical
 435 notes. BMC Med Inform Decis Mak 2018;18(Suppl 3):74.
- 436 26. Slavova S, Quesinberry D, Costich JF, et al. ICD-10-CM-Based Definitions for Emergency
- 437 Department Opioid Poisoning Surveillance: Electronic Health Record Case Confirmation
- 438 Study. Public Health Rep Wash DC 1974 2020;135(2):262–9.
- 439 27. Weiss AJ, McDermott KW, Heslin KC. Table 1, ICD-10-CM diagnosis codes defining
- different opioid-related conditions [Internet]. 2019 [cited 2021 Jan 18];Available from:
- 441 http://www.ncbi.nlm.nih.gov/books/NBK538344/table/sb247.tab1/
- 442 28. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of
- 443 Any Classifier [Internet]. 2016 [cited 2024 Oct 31];Available from:
- 444 http://arxiv.org/abs/1602.04938
- 445 29. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep
- 446 Learning Library [Internet]. 2019 [cited 2024 Oct 31];Available from:
- 447 http://arxiv.org/abs/1912.01703

It is made available under a CC-BY-NC-ND 4.0 International license .

448	30. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for
449	reporting clinical prediction models that use regression or machine learning methods. BMJ
450	2024;385:e078378.

- 451 31. Slavova S, Quesinberry D, Costich JF, et al. ICD-10-CM-Based Definitions for Emergency
- 452 Department Opioid Poisoning Surveillance: Electronic Health Record Case Confirmation
- 453 Study. Public Health Rep Wash DC 1974 2020;135(2):262–9.
- 454 32. Ranapurwala SI, Alam IZ, Pence BW, et al. Development and validation of an electronic
- 455 health records-based opioid use disorder algorithm by expert clinical adjudication among
- 456 patients with prescribed opioids. Pharmacoepidemiol Drug Saf 2023;32(5):577–85.
- 457 33. Chhabra N, Smith D, Dickinson G, et al. Trends and Disparities in Initiation of
- 458 Buprenorphine in US Emergency Departments, 2013-2022. JAMA Netw Open
- 459 2024;7(9):e2435603.

460