ODD: A Benchmark Dataset for the NLP-based Opioid Related Aberrant Behavior Detection

Anonymous ACL submission

Abstract

Opioid related aberrant behaviors (ORAB) present novel risk factors for opioid overdose. This paper introduces a novel biomedical natural language processing benchmark dataset named ODD, for ORAB Detection Dataset. ODD is an expert-annotated dataset designed to identify ORAB from patients' EHR notes and classify them into nine categories; 1) Confirmed Aberrant Behavior, 2) Suggested Aberrant Behavior, 3) Opioids, 4) Indication, 5) Diagnosed opioid dependency, 6) Benzodiazepines, 7) Medication Changes, 8) Central Nervous System-related, and 9) Social Determinants of Health. We explored two state-of-the-art natural language processing models (finetuning and prompt-tuning approaches) to identify ORAB. Experimental results show that the prompt-tuning models outperformed the finetuning models in most cateogories and the gains were especially higher among uncommon categories (Suggested aberrant behavior, Diagnosed opioid dependency and Medication change). Although the best model achieved the highest 86.92% on area under precision recall curve, uncommon classes (Suggested Aberrant Behavior, Diagnosed Opioid Dependence, and Medication Change) still have a large room for performance improvement.

1 Introduction

The opioid overdose (OOD) crisis has had a striking impact on the United States, not only threatening citizens' health (Azadfard et al., 2022) but also bringing about a substantial financial burden (Florence et al., 2021). According to a report by the Centers for Disease Control and Prevention (2023), OOD accounted for 110,236 deaths in a single year in 2022. In addition, fatal OOD and opioid use disorder (OUD) cost the United States $1.04 trillion in 2017 and that figure rose sharply to $1.5 trillion in 2021 (Beyer, 2022). Identifying patients at risk of OOD could help prevent serious consequences (Marks et al., 2021).

The opioid crisis is multifaceted, with factors like inadequate health insurance coverage (Blumenthal and Seervai, 2017), regulatory lapses (Kolodny, 2020), and profit-motivated campaigns by pharmaceutical firms (Haffajee and Mello, 2017) contributing to its complexity. Countermeasures include deploying Prescription Drug Monitoring Programs (PDMPs) (Center for Disease Control and Prevention , 2023), enhancing addiction drug education for healthcare providers (Dowell et al., 2022), and developing less addictive drugs (Thomas and Ornstein, 2017). Notably, PDMPs are data-driven systems tailored to detect patients at risk of OUD. By leveraging data analytics, these systems have successfully shielded many from critical OOD outcomes (Paulozzi et al., 2011).

Opioid-Related Aberrant Behaviors (ORABs) or Aberrant Drug Related Behaviors (ADRBs) are patient behaviors that may indicate prescription medication abuse (Fleming et al., 2008). ORABs can be categorized into confirmed aberrant behavior and suggested aberrant behavior (Portenoy, 1996; Laxmaiah Manchikanti et al., 2008; National Institute on Drug Abuse, 2023). Herein, confirmed aberrant behaviors have a clear evidence of medication abuse and addiction while suggested aberrant behaviors do not have a clear evidence (National Institute on Drug Abuse, 2023). Table 1 presents examples of such categories.

ORABs are not only clinically significant due to their strong association with OOD (Wang, 2022) and drug misuse (Maumus et al., 2020), but they also pose intriguing and challenging problems for natural language processing (NLP). This is for two primary reasons. Firstly, unlike other BioNLP tasks where reliance is primarily on medical terms or jargon (Kwon et al., 2022), ORABs encompass various behavioral patterns. These include attempts to deceive clinicians, contradictory statements, and scenarios that necessitate inference based on common sense. Secondly, given the rarity of ORABs in

ORAB Type	Example
Confirmed Aberrant Behavior	Misuse of legal substances (e.g. Alcohol)
	Falsification of prescription—forgery or alteration
	Injecting medications meant for oral use
Suggested Aberrant Behavior	Asking for or even demanding, more medication
	Asking for specific medications
	Reluctance to decrease opioid dosing once stable

Table 1: ORAB examples

patients prescribed opioids (Nadeau et al., 2021), it's crucial to consider label bias.

Previously, ORABs have been detected by monitoring opioid administration (e.g., frequency and dosage) (Rough et al., 2019) or self-reported questionnaires (Adams et al., 2004; Webster and Webster, 2005). However such measurements do not include the full spectrum of ORABs (e.g., medication sharing, denying medication changing). In addition, patients can obtain opioids from multiple resources (e.g. illegal purchase and medication sharing), which are not captured in the structured data. It has been known that ORABs are widely described in EHR notes and natural language processing (NLP) techniques can be used to identify ORABs (Lingeman et al., 2017). Nonetheless, the previous study relied on a small amount of annotated notes, which were not publicly available. Moreover, the previous work only considered ORABs as a binary classification (present or not) and only explored traditional machine learning models (e.g., support vector machine (SVM)).

This paper proposes ORAB detection that is a novel Biomedical NLP (BioNLP) task. We also introduce an **O**RAB **D**etection **D**ataset **(ODD)** which is *large-size*, *expert-annotated*, and *multi-label classification* benchmark dataset corresponding to the task. For this, we first designed a robust and comprehensive annotation guideline that labels text into nine categories which encompass two types of ORABs (Confirmed Aberrant Behavior and Suggested Aberrant Behavior) and seven types of auxiliary opioid-related information (Opioids, Indication, Diagnosed Opioid Dependency, Benzodiazepines, Medication Change, Central Nervous System Related, Social Determinant of Health). Using the guideline, domain experts annotated 750 sampled EHR notes of 500 opioid-treated patients extracted from MIMIC-IV database (Johnson et al., 2021). Overall, we found 399 EHR notes with a opioid prescription. Overall, we annotated 3,718 instances with 162 ORABs instances (115 for Confirmed Aberrant Behavior and 47 for Suggested Aberrant Behavior) on 2,840 sentences.

Experiments conducted on two ORAB detection models based on state-of-the-art (SOTA) natural language processing (NLP) models; traditional finetuning (Devlin et al., 2018) and prompt-based tuning (Webson and Pavlick, 2022) approaches. The experimental results on MIMIC showed that prompt-based tuning models surpass finetuning models in almost all categories (eight out of nine). When the numbers of instances were less than 100 ($uncommon\ categories$: Suggest Aberrant Behavior, Diagnosed Opioid Dependency, and Medication Change), the performance improvement was greater, in particular, the Medication Change and Suggest Aberrant Behavior classes achieve performance improvements of over 7%p and 13%p respectively. ODD will be published after being accepted.

The main contributions of this paper can be organized as follows:

- This paper introduces a new Biomedical NLP (BioNLP) task **ORAB detection** for extracting information related to a patient's risk of opioid addiction and abuse from EHR notes. We also curate a corresponding benchmark dataset, named **ODD**, an expert-annotated dataset for the ORAB detection task.

- We present the experimental results of two state-of-the-art NLP models as baseline performances for the benchmark dataset. Moreover, we report comprehensive data and error analyses to guide future studies in constructing improved models.

2 Related Work

NLP-based Opioid Abuse Analysis Recently, with the development of NLP technology, studies have been actively conducted to analyze information relevant to opioid abuse and OOD from text (e.g. EHR notes, social media) (Sarker et al., 2019; Blackley et al., 2020; Goodman-Meza et al., 2022; Zhu et al., 2022; Singleton et al., 2023). Studies have explored a broad range of NLP techniques to identify OUD (Zhu et al., 2022). Zhu et al. (2022) developed a keyword-based OUD detection model for patients who have been treated with chronic opioid therapy. Their NLP models were able to uncover OUD cases that would be missed using the International Classification of Diseases (ICD) codes alone. Singleton et al. (2023) proposed a multiple-phase OUD detection approach using a combination of dictionary and rule-based approaches. Blackley et al. (2020) developed feature engineering-based machine learning models. Herein, the authors demonstrated that the machine learning models outperformed a rule-based one that utilizes keywords.

Other works adopted NLP to study factors associated with opioid abuse. Goodman-Meza et al. (2022) utilized text features such as term frequency–inverse document frequency (TF-IDF), concept unique identifier (CUI) embeddings, and word embeddings to analyze substances that contribute to opioid overdose deaths. Sarker et al. (2019) conducted a geospatial and temporal analysis of opioid-related mentions in Twitter posts. They found a positive correlation between the rate of opioid abuse-indicating posts and opioid misuse rates and county-level overdose death rates.

The ORAB detection task is similar to the studies above in that it analyzes drug abuse-related information using NLP approaches. However, different from the previous studies that mainly depend on keywords such as drug mentioning, the ORAB detection is a more challenging NLP task considering that it needs to identify various and complex linguistic patterns such as trying to deceive physicians (Passik and Kirsh, 2007) and emotional reaction on opioid prescription (Lingeman et al., 2017).

ORAB Risk Assessment and Detection Webster and Webster (2005) introduced a risk management tool that monitors ORABs by scoring a patient's self-reports on risk factors (history of family and personal substance abuse, history of preadolescent sexual abuse, and psychological illness) related to substance abuse. Then, each patient is categorized into three risk levels (low risk, moderate risk, and high risk) according to the sum of the scores. Other studies (Schloff et al., 2004; Sullivan et al., 2010; Katz et al., 2010; Tudor, 2013; Rough et al., 2019) suggest detecting ORAB by relying on diagnostic criteria based on structured information such as the frequency of opioid dosage, the number of opioid prescribers, and the number of pharmacies. Although the above methodologies can detect patients at risk of ORABs with high precision, the recall was low (Rough et al., 2019).

The most relevant work is Lingeman et al. (2017). However, as described earlier, Lingeman et al. (2017)'s work relied on a small scaled EHR notes which is not publicly available. In contrast, ODD consists of a larger dataset which is publicly available. Furthermore, ODD's annotation scheme provides rich sub-categorized aberrant behaviors (suggested and confirmed) and additional opioid-related information. In contrast, Lingeman et al. (2017)'s study was designed as a binary classification task to detect ORABs. Finally, we leverage the SOTA deep learning models that the previous work Lingeman et al. (2017) did not explore.

3 ORAB Detection Dataset

3.1 Data Collection

The source of the first dataset is made up of publicly available fully de-identified EHR notes of the MIMIC-IV (Johnson et al., 2023). ORABs are uncommon events. To increase the likelihood that our annotated data incorporate ORABs, we sorted out patients at risk of opioid misuse based on repetitive opioid use and diagnosis related to opioid misuse. Specifically, we first extracted EHR notes mentioning opioids with the generic and brand name of opioid medications. In addition, we selected patients diagnosed based on their ICD codes. Detailed information on opioid medications (and their generic names), and ICD codes utilized for filtering EHR notes are presented in Appendix A.

Among 331,794 EHR notes of 299,712 patients in MIMIC-IV database, we found that approximately 57% of patients were prescribed opioids during their hospitalization. Then, we selected patients who were repeatedly prescribed (more than twice) opioids. In addition, we chose patients who were diagnosed with drug poisoning and drug dependence based on the ICD codes. Overall, there are 3,904 patients who are satisfied the aforementioned conditions. Among them, we randomly select 750 notes from a randomly sampled 500 patients for annotation.

3.2 Data Annotation

For the annotation process, we initially identified nine categories, which include two ORABs (confirmed aberrant behavior and suggested aberrant behavior), as well as seven additional pieces of information relevant to opioid usage and misuse. These categories are briefly outlined in Table 2. The annotation process was iterative, with continuous refinement of the EHR note annotations and annotation guidelines. An interdisciplinary team of addiction medicine, biostatisticians, and NLP specialists collaboratively discussed and developed these guidelines. This rigorous approach yielded a comprehensive annotation guideline adept at addressing language variations and ambiguities in clinical

Category	Definition	Example
Confirmed Aberrant Behavior	Evidence confirming the loss of control of opioid use, specifically aberrant usage of opioid medications.	[Patient] admits that he has been sharing his Percocet with his wife, and that is why he has run out early.
Suggested Aberrant Behavior	Evidence suggesting loss of control of opioid use or compulsive/inappropriate use of opioids.	[Patient] states that 'that [drug] won't work; only [X drug] will and I won't take any other'
Opioids	The mention or listing of the name(s) of the opioid medication(s) that the patient is currently prescribed or has just been newly prescribed.	Oxycodone has been known to make [the patient] sleepy at 5 mg.
Indication	Patients are using opioids under instructions.	[The patient] is in a daze.
Diagnosed Opioid Dependency	Patients have the condition of being dependent on opioids, have chronic opioid use, or is undergoing opioid titration	[The patient] is in severe pain and has been taking [opioid drug] for [time].[HY1]
Benzodiazepines	Patients are co-prescribed benzodiazepines.	Valium has been listed in patient medication list.
Medicine Changes	Change in opioid medicine, dosage, and prescription since the last visit.	[Patient] reports that his previous PCP just recently changed his pain regimen, adding oxycodone.
Central Nervous System Related	CNS-related terms/terms suggesting altered sensorium.	[Patient] reported to have nausea after taking [drug].
Social Determinants of Health	The nonmedical factors that influence health outcomes	[Patient] divorced a years ago.

Table 2: The definitions and examples of the categories of ODD.

Socio-demographic type	Group	# of patients (percentage)
Gender	Male	168 (51.69%)
	Female	157 (48.31%)
Age	19-25	14 (4.31%)
	26-35	34 (10.46%)
	36-45	59 (18.15%)
	46-55	80 (24.62%)
	56-65	69 (21.23%)
	66-75	40 (12.31%)
	> 75	29 (8.92%)
Total		325 (100%)

Table 3: Socio-demographic statistics of the cohort.

Categories	Instances
Confirmed Aberrant Behavior	115 (3.09%)
Suggested Aberrant Behavior	47 (1.26%)
Opioids	1,678 (45.13%)
Indication	558 (15.01%)
Diagnosed Opioid Dependency	67 (1.80%)
Benzodiazepines	417 (11.22%)
Medication Change	139 (3.74%)
Central Nervous System Related	542 (14.58%)
Social Determinants of Health	155 (4.17%)
Total	3,718 (100%)

Table 4: Categorical distribution of the annotated instances.

narratives related to opioid misuse. For detailed descriptions of the categories, please refer to Appendix A.2. The annotation guidelines developed can be accessed in the 'annotation_guideline.pdf' file available in the supplementary data.

EHR notes were annotated independently by two domain experts who are familiar with medical literature and EHR notes by following the annotation guidelines. Herein, the primary annotator [1] annotated all EHR notes with eHOST (eHOST, 2011) annotation tool. The other annotator [2] coded 25 of the EHRs of the primary annotator with the same environment to compute inter-rater reliability with Cohen's kappa (Warrens, 2015). As a result, the inter-rater reliability shows strong agreement ($\kappa = 0.87$) between the annotators. After annotation, among 750 notes, we could find 399 notes of 325 patients who are current opioid prescription. The socio-demographic statistics on the final patient cohort can be found in Table 3. Overall, there are 2,840 sentences that contain explicit evidences at least one of the target categories.

3.3 Annotation Statistics

Table 3 shows the statistics of the annotated instances from the 2,840 sentences. Herein, MIMIC dataset consist of 3,718 instances annotated from the EHRs. Especially, we can notice that 'confirmed aberrant behavior' and 'suggested aberrant behavior' in EHRs are relatively rare events only accounting for 162 (4.25%); 115 (3.09%) for confirmed aberrant behavior and 47 (1.26%) for suggested aberrant behavior. The 'Opioids,' 'Indication,' and 'Central nervous system related' are majority classes accounting for over 74% of overall instances while the other categories are around or less than 10% each.

[1]A master of public health

[2]A medical doctor affiliated with the addiction medicine

4 Task Definition and Evaluation Criteria

Task Definition The ORAB detection is an **information extraction task** that identifies whether an input text contains ORABs (Confirmed, and Suggested aberrant behaviors) and information relevant to opioid usage. In addition, since all labels can be co-occurred together in a sentence, we formulate the **multi-label classification**.

Evaluation Criteria Previous study on NLP-based ORAB detection (Lingeman et al., 2017) utilizes accuracy as an evaluation criterion. However, since the labels in the dataset are highly imbalanced (in Table 4), the accuracy may mislead performance on rare classes since it can overestimate true negative cases (Bekkar et al., 2013). Thus, as main evaluation criteria, we adopt the Area Under Precision-Recall Curve (AUPRC) and the F1-score that have widely utilized for the performance evaluation of the binary classifiers on highly biased labels (Ozenne et al., 2015).

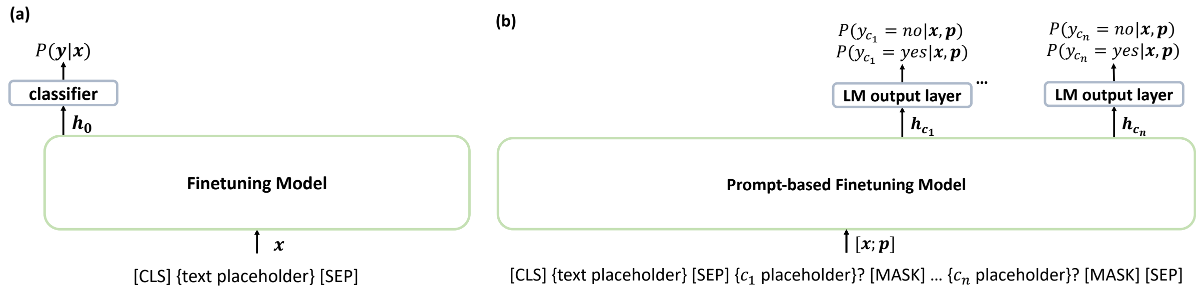

Figure 1: The figures illustrate the conceptual architectures of our ORAB detection models. (a) demonstrates a finetuning model and (b) depicts a promtuning model. Herein, $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{p}$ indicate input text, output labels, and prompt text respectively. $\mathbf{h_i}$ is the hidden vector representation of the i^{th} input token. EHR text input to '{text placeholder}'. The name of each category ($c_{1...n}$) in Table 2 is input at '{$c_{1...n}$ placeholder}'.

5 ORAB Detection Models

This section demonstrates pretrained Language Model (LM) based ORAB detections models; traditional fine-tuning model (Zahera et al., 2019) and prompt-tuning model. The prompt-based finetuning model has shown advantages in rare category classification (e.g. zero-shot or few-shot classification) (Yang et al., 2023). Figure 1 demonstrates the baseline ORAB detection models.

5.1 Finetuning Models

The most common way to construct classification models using a pretrained language model (LM) is to employ finetuning, as illustrated in Figure 1(a). In this approach, the input text $\mathbf{x}$ is passed through the fine-tuning model. The hidden representation vector of the first token '[CLS]' ($\mathbf{h_0}$) is then used as input for the classifier. Here, W_c and $\mathbf{b_c}$ represent the weight matrix and bias, respectively. The classifier calculates the probability distribution over output labels $\mathbf{y}$ using the sigmoid function.

5.2 Prompt-based Finetuning Models

Although finetuning on pretrained LMs has been successfully applied to most of NLP tasks (Devlin et al., 2018), it is still known that finetuning still requires considerable annotated examples to achieve a high performance (Webson and Pavlick, 2022; Yang et al., 2023). Thus, uncommon categories in ODD may be a performance bottleneck.

The widely recognized technique of prompt-based finetuning, as demonstrated in studies by Gao et al. (2021) and Yang et al. (2022), utilizes a template to transform a downstream task into a language modeling problem by incorporating masked language modeling and a predefined set of label words, effectively enabling effective few-shot learning capabilities.

We utilize the full name of each class to curate the prompt text p. Specifically, the prompts for each class are arranged in the same order as Table 1, following the template "[c_i placeholder]? [MASK]" where c_i represents the name of the i^{th} class. The prompt text is then concatenated with $\mathbf{x}$, distinguished by a separator token "[SEP]," and fed into a prompt-based tuning model. Next, we calculate the probability that the language model (LM) output of the masked token corresponding to each class would be a positive word or a negative word. Following the approach of Gao et al. (2021), we define the positive word as 'yes' and the negative word as 'no'. Thus, the probability of 'yes' for the i^{th} class c_i ($P(y_{c_i} =$'yes'$|\mathbf{x}, \mathbf{p})$) can be interpreted as the probability that c_i is included in the input text $\mathbf{x}$, and vice versa.

6 Experiment

6.1 Experimental Environment

Experimental Models To verify the generalizability of experimental results, we prepared two different LMs pretrained on Biomedical literacy; BioBERT (Lee et al., 2020) and BioClinicalBERT (Alsentzer et al., 2019) Herein, 'Finetune' and 'Prompt' indicate an LM trained on ODD via finetuning (in Section 5.1) and prompt-based finetuning (in Section 5.2) respectively.

Experimental Setting For the experiments, we conducted 5-fold cross-validation and report the average performance and standard deviation. We adopted a loss function as a binary cross entropy for finetuning models and categorical cross entropy for prompt-based finetuning models (LeCun et al., 2015). Moreover, we selected the optimizer as AdamW (Loshchilov and Hutter, 2017).

Categories	Fine-tuning				Prompt-based Fine-tuning			
	BioBERT		BioClinicalBERT		BERT		BioClinicalBERT	
	AUPRC	F1	AUPRC	F1	AUPRC	F1	AUPRC	F1
Confirmed Aberrant Behaviors	80.76±8.96	60.39±9.92	81.31±7.48	64.74±5.74	**84.18±6.29**	**68.23±7.85**	**88.11±7.94**	**72.25±8.50**
Suggested Aberrant Behaviors	30.06±5.43	22.52±11.60	33.57±4.61	32.73±5.40	**50.04±15.05**	**44.54±15.31**	**46.88±11.23**	**49.70±10.67**
Opioids	98.85±0.62	96.96±0.60	98.73±0.39	97.48±0.19	**99.44±0.20**	**97.46±0.42**	**99.52±0.20**	**97.65±0.53**
Indication	97.00±1.73	93.42±2.15	97.28±1.56	**94.09±2.58**	**97.79±0.85**	**93.96±1.82**	**97.77±0.78**	93.37±1.60
Diagnosed Opioid Dependency	64.42±25.87	54.25±18.65	75.67±15.41	59.37±9.73	**82.80±14.69**	**71.88±13.83**	**84.20±7.58**	**70.23±15.04**
Benzodiazepines	**97.27±2.04**	93.99±1.42	**97.19±1.23**	96.36±1.39	96.56±2.56	**96.14±0.97**	96.68±1.11	**97.31±0.51**
Medication Change	56.26±17.67	50.28±11.01	68.13±5.40	63.17±6.72	**76.82±4.45**	**68.36±5.92**	**75.25±4.35**	**67.20±4.79**
Central nervous system related	98.09±0.79	89.85±3.02	97.97±0.70	89.33±2.33	**98.32±1.20**	**94.65±1.13**	**98.18±0.67**	**90.23±1.49**
Social Determinants of Health	88.57±6.24	79.33±8.93	93.09±5.30	85.53±9.15	**94.82±2.62**	**91.49±3.27**	**95.68±1.76**	**91.45±2.63**
Macro Average	79.03±24.07	71.22±25.74	82.55±21.50	75.87±22.11	**86.75±15.99**	**80.75±18.37**	**86.92±17.04**	**81.04±16.78**

Table 5: This table presents the experimental results of ODD on BioClinicalBERT and BioBERT. Note that, 'Finetune' and 'Prompt' indicate models are trained with the finetuning and prompt-based finetuning respectively. Each value stands for the average and the standard deviation of five-fold cross-validation results and average scores with higher values are bolded. Finally, * stands for the statistical significance ($p < .05$) of performance improvement between fine-tuning results and prompt-based fine-tuning results.

Hyper-parameter Setting We conducted the grid search with the following range of possible values for each hyper-parameter: {2e-5, 3e-5, 5e-5} for learning rate, {4, 8, 16} for batch size, {2,3,4} for the number of epoch. Herein, we choose the hyper-parameters that achieved the best performance on the first fold of the BioClinicalBERT finetune environment with the grid search. Finally, we chose 3e-5 for learning rate, 8 for batch size, and 3 for the number of epochs.

Others To evaluation the statistical significance in performance between models, we adopted student's t-test (Student, 1908). In all of the experiments, we keep the random seed as 0. Finally, all experiments were performed on an NVIDIA P40 GPU with CentOS 7 version.

6.2 Experimental Results

Table 5 shows the experimental results of the five-fold cross-validation on the experimental models. To sum up, the performance range shows [79.03-86.92] based on the macro average AUPRC and [71.22-81.04] based on the macro average F1. Especially, prompt-based finetuning models outperformed the finetuning models in both BioClinicalBERT and BioBERT with large margins of 4.37%p and 7.72%p in AUPRC respectively. Herein, BioClinicalBERT-based models achieved a higher performance compared to BioBERT-based models. These results are not surprising because the pre-training BioClinicalBERT's corpora contain EHR notes from MIMIC-III (Johnson et al., 2016) that is the previous version of our target database MIMIC-IV and both databases were collected from the same hospital.

Otherwise, the performance among the classes has a large spectrum. For example, in the BioClinicalBERT finetuned model, the class with the highest performance (Opioids) is 98.73 in AUPRC, which is more than double the performance gap compared to 33.57 of the lowest class (Suggested Aberrant Behaviors). Herein, the performance gap between these classes is related to the number of instances. For example, the dominant classes, Opioids, Indication, Benzodiazepines, and Central Nervous System Related show very high performance with scores of 98.73, 97.28, 97.19 and 97.97, respectively. However, it can be seen that the detection performance of the uncommon categories is inferior showing 33.57 for Suggested Aberrant Behavior, 75.67 for Diagnosed Opioid Dependency, and 68.13 for Medication Change. Moreover, we can notice that the performance results show the same trend in BioBERT.

Overall, prompt-based finetuning contributes to enhanced performance in nearly all environments (16 out of 18 cases), with the sole exception being Benzodiazepines on BioClinicalBERT, where the performance difference was negligible (-0.51%p). The introduction of prompt-based finetuning resulted in significant improvements, particularly in uncommon categories. The performance of prompt-based finetuning on BioClinicalBERT and BioBERT increased by 13.31%p and 19.98%p respectively in the Suggested Aberrant Behavior class. In the Diagnosed Opioid Dependence class, the performance of prompt-based finetuning on BioClinicalBERT and BioBERT improved by 8.53%p and 18.83%p, respectively. Lastly, in the Medication Change class, the performance saw a rise of more than 20%p on BioBERT. Despite these advancements, further performance improvements are still needed for uncommon categories.

7 Discussion

7.1 Error Analysis

First of all, we demonstrate that quantitative aspect of errors. For this, we gathered all the results of

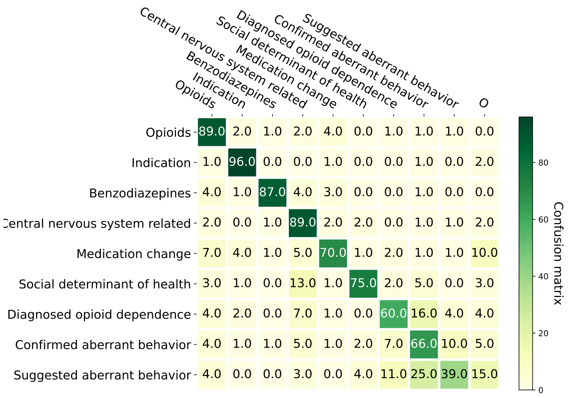

Figure 2: A multi-label confusion matrix among categories. Herein, 'O' indicates the none of any categories.

the test sets of 5-fold cross validation then calculated a normalized multi-label confusion matrix (Heydarian et al., 2022). Figure 2 shows that there are confusions between two specific classes: confirmed aberrant behavior and suggested aberrant behavior. The confusion rates were found to be 10.0% and 25.0%, respectively, for these classes. This indicates that the confirmed and suggested aberrant behaviors were the classes most prone to being mistaken for one another in our test sets. In addition, there are large confusions among diagnosed opioid dependence, confirmed and suggested aberrant behaviors which is 20% in total.

We also report the qualitative aspect of errors by scrutinizing the first fold of the BioClinicalBERT prompt-based finetuning model. Especially, this paper focuses on error cases of the three uncommon categories: Suggested Aberrant Behavior, Diagnosed Opioid Dependency, and Medication Change.

Firstly, regarding Suggested Aberrant Behavior, we identified a problem with insufficient data on specific abnormal behavior patterns. For instance, consider the sentence "He is requesting IV morphine for his chest pain." This is a clear example of suggested aberrant behavior as the patient is asking for a specific medication (IV morphine). However, due to a lack of similar pattern sentences in the data, the model finds it challenging to learn these patterns.

Likewise, in the case of Medication Change, the sentence "The only exception being that his home dilaudid 4mg was increased from every 6h to every 4h" represents a medication change due to alterations in the drug administration time interval. In this instance, the ML model might overlook the significance of the change in time intervals due to the scarcity of similar patterns.

Categories	Age		Gender	
	age$<$45	age $\geq$ 45	Female	Male
	AUPRC	AUPRC	AUPRC	AUPRC
Confirmed Aberrant Behaviors	**94.38±5.51**	85.86±10.01	89.29±8.46	89.15±8.55
Suggested Aberrant Behaviors	**58.39±11.32**	40.84±27.43	51.46±10.01	51.10±24.64

Table 6: Experimental results on different age and gender groups.

Furthermore, when dealing with Diagnosed Opioid Dependency, we noticed that a model heavily relies on specific keywords. For example, the sentence "Insulin Dependent DM c/b has peripheral neuropathy..." was classified as opioid dependence, which is a misclassification. This error occurred due to the reliance on the keyword 'dependent', despite the fact that insulin is not an opioid.

Finally, we observed some text requires commonsense to correctly predict the label. For example, the text "3 pitcher sized cocktail daily," indicates the patient is addicted to alcohol which is a confirmed aberrant behavior. However, the prediction probability for this sentence is 0.31%, so the training model totally fails to identify that this sentence stands for confirmed aberrant behavior. This is because, different from other examples where keywords such as alcohol addiction and alcohol abuse are presented, in order to understand the above example, it is understood that the 3 pitcher cocktail is an excessive dose and daily consumption is clear evidence of alcohol abuse.

7.2 Socio-demographic Analysis

Patient groups with varying socio-demographics frequently exhibit distinct characteristics. To examine the disparities among these groups, we carried out studies that disaggregated the data based on two socio-demographic factors (age and gender) in Table 6.

Gender The gender of the patients has little effect on the aberrant behavior detection performance, which means that the bias between genders is trivial. In fact, the male and female groups account for almost the same proportion of the total number of patients.

Age We divided patients into two groups based on age 45, which is the standard for specifying the risk according to the patient's age (Brott et al., 2020), and evaluated performance of aberrant behaviors. Experimental results showed that the performances of aberrant behaviors are significantly different between two age groups. Especially, the performance of the younger age group achieved

Confirmed Aberrant Behaviors		
Subcategories	age $<$ 45	age $\geq$ 45
Self-escalating dose	0	6
Using opioids outside of the prescriber's purpose	1	3
Substance abut OTHER than prescription opioids	1	2
Evidence of a patient selling or giving opioids to others	0	1
Suggested Aberrant Behaviors		
Subcategories	age $<$ 45	age $\geq$ 45
Clinician's concern on opioids	2	1
Obtaining opioids from non-medical sources	0	2
Patient's request for a higher or specific opioid	3	2
Obtaining opioids from multiple-medical sources	1	2
Patient's strong emotion/opinion on opiods	0	1
Others	1	1

Table 7: Subcategorical error analysis on different age groups.

Categories	BioClinicalBERT		T5 Paraphrasing	
	AUPRC	F1	AUPRC	F1
CAB	88.11 $\pm$ 7.94	72.25$\pm$8.50	91.03 $\pm$ 6.29	85.29 $\pm$ 6.26
SAB	46.88 $\pm$ 11.23	49.70 $\pm$ 10.67	62.66 $\pm$ 14.71	53.38 $\pm$ 13.17

Table 8: Experimental results of the data augmentation with the LLM paraphrasing on confirmed aberrant behaviors (CAB) and suggested aberrant behaviors (SAB).

higher performance although the proportion of patients in the older group is greater (over 45: 69.23%, less than 45: 30.77%).

We speculate that this is because more diverse patterns of aberrant behaviors are observed in the older group. Table 7 shows the error analysis results for each age group. We can see that both confirmed aberrant behaviors and suggested aberrant behaviors in the older group show more diverse aberrant behavior patterns than in the younger group.

7.3 Potential Application of LLMs

One prospective application of LLMs on this task is data augmentation. For example, we additionally conducted data augmentation experiments with a LLM, Flan T5 XL (Chung et al., 2022), for data augmentation with a simple prompt.

"Rewrite: {input text holder}"

Here, we generated three paraphrased sentences for all sentences of the train set of each fold and add them to the training set. Experimental results showed that the data augmentation helps to enhance the performance of aberrant behavior detection at BioClinicalBERT + Prompt-based environment.

The results in Table 8 demonstrate that data augmentation could be a promising solution for this task. Especially the performance on one of the uncommon classes "diagnosed opioid dependence" increased significantly. However, due to the various linguistic patterns of suggested aberrant behaviors, there is still room for performance improvement by paraphrasing alone although the performance enhanced significantly. Through developed data augmentation method with LLMs in the future, we can expect additional performance improvements in suggested aberrant behaviors and medication change classes. Entire experimental results containing additional categories can be found in Appendix B.

7.4 Merits & Demerits

Our research can have the following positive impacts. Firstly, the information extracted by ORAB detection models can be utilized for various studies and systems aimed at addressing opioid abuse. For instance, since ORABs serve as important evidence of OUD, they can be used as key features in opioid risk monitoring systems. Additionally, this information can be leveraged to detect a patient's risk of OOD or opioid addiction at an earlier stage, thereby assisting in the prevention of fatal OOD cases. Consequently, by supporting efforts to mitigate future opioid overdoses, our research would contribute to maintaining people's health.

However, it is important to acknowledge that our work may have certain negative social impacts. As previously mentioned, ORAB detection can be utilized to strengthen opioid monitoring systems, but this may unintentionally encroach upon the autonomy of doctors (Clark et al., 2012). Indeed, in previous studies, although strict opioid prescription policies and prescription drug monitoring programs (PDMPs) help patients forestall opioid misuse or overuse (McCauley et al., 2016; Dowell et al., 2016), oligonalgesia (Dowell et al., 2016), has been pointed out as a possible side effect of PDMPs (Cantrill et al., 2012).

8 Conclusion

This paper introduces a novel BioNLP task called ORAB detection, which aims to identify two ORAB categories and seven categories relevant to opioid usage from EHR notes. We also present the associated benchmark dataset, ODD. The paper provides baseline models and their performances on ODD. To this end, we trained two SOTA pretrained LMs using a fine-tuning approach and prompt-based fine-tuning. Experimental results demonstrate that the performance in three uncommon categories was notably lower compared to the other categories. However, we also discovered that prompt-based fine-tuning can help mitigate this issue. Additionally, we provide various error analysis results to guide future studies.

Ethical Consideration

First, one prospective concern is whether is it legal to screen patients and provide prior medical history without their consent. According to the U.S. Department of Health and Human Service (2021), "The Health Insurance Portability and Accountability Act (HIPAA) regulation allows health care providers to disclose protected health information about an individual, without the individual's authorization, to another health care provider for that provider's treatment of the individual" (§ 45 CFR 164.506). Health care providers can be defined at §45 CFR PART 171 (The Office of the National Coordinator for Health Information Technology, 2020):

- hospital, skilled nursing facility, nursing facility, home health entity or other long-term care facility, health care clinic, community mental health center, renal dialysis facility, blood center, ambulatory surgical center, emergency medical services provider, Federally qualified health center, group practice, a pharmacist, a pharmacy, a laboratory, a physician, a practitioner, a provider operated by, or under contract with, the Indian Health Service or by an Indian tribe, tribal organization, or urban Indian organization, a rural health clinic, a covered entity under section 256b of this title, an ambulatory surgical center, a therapist, and any other category of health care facility, entity, practitioner, or clinician determined appropriate by the Secretary.

Another consideration is the dataset's quality. We attempted to ameliorate this issue by developing a thoroughly systematic annotation guideline. First of all, we used an iterative process throughout the annotation, going back and forth between EHR note annotations and establishing annotation guidelines. The guidelines were discussed among an interdisciplinary team of experts in addiction (3), biostatisticians (2), and NLP (2). In this process, we curated a comprehensive annotation guideline, which addresses various aspects of how to handle language variations and ambiguities in clinical narratives related to this annotation task.

In addition, the data annotation quality might be a concerned since it requires specialized medical knowledge. Although the main annotator's annotations are almost perfectly aligned with the domain expert ($\kappa = 0.87$), it is still a question whether the primary annotator is consistent. Thus, to analyze annotation quality, the primary annotator performed re-annotation on 25 sampled notes. At this time, initial annotation was performed on April 21-May 26, and re-annotation was performed on August 25-26, about 3 months later. Results The Kappa score of the two annotations was $\kappa = 0.96$, which was almost perfectly consistent with the previous annotations. This implies that the annotation of the dataset used in this paper is consistent and reliable.

Limitation & Future Work

The ORAB detection task relies on EHR notes. Thus, if health providers do not recognize the patient's abnormal signs, they may not describe aberrant behaviors in a note. In this case, our approach cannot detect ORABs. In the future, we will develop an algorithm that detects a wider spectrum of ORABs by combining them with previous structured information-based methods.

Another limitation is that our data source was derived from a single hospital's EHR database. Although many existing studies have been conducted based on the MIMIC database, this does not guarantee that the system developed as a result of this study can be migrated to different clinical settings. Therefore, we plan to perform annotation based on annotation guidelines in additional clinical environments in the future and evaluate the model's performance.

Moreover, ORAB detection models still have limited performance in the uncommon categories. It is necessary to improve performance through advanced NLP approaches data augmentation (Wei and Zou, 2019), medical knowledge injection (Yang et al., 2022), or leveraging knowledge extracted from large language models (Kwon et al., 2023).

Finally, errors can cause negative downstream effects. In particular, the most significant negative downstream impact is that some errors for example misprediction of opioid dependence can lead to a false stigma to the patient which is known as one of the unintended harms of PDMPs (Haines et al., 2022).

References

Laura L Adams, Robert J Gatchel, Richard C Robinson, Peter Polatin, Noor Gajraj, Martin Deschner, and Carl Noe. 2004. Development of a self-report screening instrument for assessing potential opioid medication misuse in chronic pain patients. *Journal of pain and symptom management*, 27(5):440–459.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Mohammadreza Azadfard, Martin R. Huecker, and James M. Leaming. 2022. Opioid addiction. In *StatPearls [Internet]*.

Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. 2013. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10).

Don Beyer. 2022. The economic toll of the opioid crisis reached nearly $1.5 trillion in 2020. shorturl.at/kovz3. Accessed: 2023-03-08.

Suzanne V Blackley, Erin MacPhaul, Bianca Martin, Wenyu Song, Joji Suzuki, and Li Zhou. 2020. Using natural language processing and machine learning to identify hospitalized patients with opioid use disorder. In *AMIA Annual Symposium Proceedings*, volume 2020, page 233. American Medical Informatics Association.

David Blumenthal and Shanoor Seervai. 2017. To combat the opioid epidemic we must be honest about all its causes. *Harvard Business Review*.

Nathan R Brott, Elisha Peterson, and Marco Cascella. 2020. Opioid, risk tool. *StatPearls*.

Stephen V Cantrill, Michael D Brown, Russell J Carlisle, Kathleen A Delaney, Daniel P Hays, Lewis S Nelson, Robert E O'Connor, AnnMarie Papa, Karl A Sporer, Knox H Todd, et al. 2012. Clinical policy: critical issues in the prescribing of opioids for adult patients in the emergency department. *Annals of emergency medicine*, 60(4):499–525.

Center for Disease Control and Prevention . 2023. Prescription drug monitoring programs (pdmps). https://www.cdc.gov/opioids/healthcare-professionals/pdmps.html. Accessed: 2023-09-24.

Centers for Disease Control and Prevention. 2023. Provisional drug overdose death counts. https://www.ncbi.nlm.nih.gov/books/NBK448203/. Accessed: 2023-03-08.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Thomas Clark, John Eadie, Peter Kreiner, and Gail Strickler. 2012. Prescription drug monitoring programs: an assessment of the evidence for best practices. *The Prescription Drug Monitoring Program Center of Excellence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Deborah Dowell, Kathleen R Ragan, Christopher M Jones, Grant T Baldwin, and Roger Chou. 2022. Cdc clinical practice guideline for prescribing opioids for pain—united states, 2022. *MMWR Recommendations and Reports*, 71(3):1.

Deborah Dowell, Kun Zhang, Rita K Noonan, and Jason M Hockenberry. 2016. Mandatory provider review and pain clinic laws reduce the amounts of opioids prescribed and overdose death rates. *Health Affairs*, 35(10):1876–1883.

eHOST. 2011. ehost: The extensible human oracle suite of tools. https://code.google.com/archive/p/ehost/. Accessed: 2023-4-10.

Michael F Fleming, James Davis, and Steven D Passik. 2008. Reported lifetime aberrant drug-taking behaviors are predictive of current substance use and mental health problems in primary care patients. *Pain Medicine*, 9(8):1098–1106.

Curtis Florence, Feijun Luo, and Ketra Rice. 2021. The economic burden of opioid use disorder and fatal opioid overdose in the united states, 2017. *Drug and alcohol dependence*, 218:108350.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pages 3816–3830. Association for Computational Linguistics (ACL).

David Goodman-Meza, Chelsea L Shover, Jesus A Medina, Amber B Tang, Steven Shoptaw, and Alex AT Bui. 2022. Development and validation of machine models using natural language processing to classify substances involved in overdose deaths. *JAMA Network Open*, 5(8):e2225593–e2225593.

Rebecca L Haffajee and Michelle M Mello. 2017. Drug companies' liability for the opioid epidemic. *New England Journal of Medicine*, 377(24):2301–2305.

Sarah Haines, Ashley Lam, Michael Savic, and Adrian Carter. 2022. Patient experiences of prescription drug monitoring programs: a qualitative analysis from an australian pharmaceutical helpline. *International Journal of Drug Policy*, 109:103847.

Mohammadreza Heydarian, Thomas E Doyle, and Reza Samavi. 2022. Mlcm: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and S Horng IV. 2021. Mimic-iv-ed. *PhysioNet*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Nathaniel Katz, Lee Panas, MeeLee Kim, Adele D Audet, Arnold Bilansky, John Eadie, Peter Kreiner, Florence C Paillard, Cindy Thomas, and Grant Carrow. 2010. Usefulness of prescription monitoring programs for surveillance—analysis of schedule ii opioid prescription data in massachusetts, 1996–2006. *Pharmacoepidemiology and drug safety*, 19(2):115–123.

Andrew Kolodny. 2020. How fda failures contributed to the opioid crisis. *AMA Journal of Ethics*, 22(8):743–750.

Sunjae Kwon, Rishabh Garodia, Minhwa Lee, Zhichao Yang, and Hong Yu. 2023. Vision meets definitions: Unsupervised visual word sense disambiguation incorporating gloss information. *arXiv preprint arXiv:2305.01788*.

Sunjae Kwon, Zonghai Yao, Harmon Jordan, David Levy, Brian Corner, and Hong Yu. 2022. MedJEx: A medical jargon extraction model with Wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

MD Laxmaiah Manchikanti, MD Sairam Atluri, and MD Andrea Trescot. 2008. Monitoring opioid adherence in chronic pain patients: Tools, techniques and utility. *Pain Physician*, 11:1.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jesse M Lingeman, Priscilla Wang, William Becker, and Hong Yu. 2017. Detecting opioid-related aberrant behavior using natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1179. American Medical Informatics Association.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Charles Marks, Daniela Abramovitz, Christl A Donnelly, Gabriel Carrasco-Escobar, Rocío Carrasco-Hernández, Daniel Ciccarone, Arturo González-Izquierdo, Natasha K Martin, Steffanie A Strathdee, Davey M Smith, et al. 2021. Identifying counties at risk of high overdose mortality burden during the emerging fentanyl epidemic in the usa: a predictive statistical modelling study. *The Lancet Public Health*, 6(10):e720–e728.

Marianne Maumus, Renée Mancini, Daniel M Zumsteg, and Dileep K Mandali. 2020. Aberrant drug-related behavior monitoring. *The Ochsner Journal*, 20(4):358.

Jenna L McCauley, J Madison Hyer, V Ramesh Ramakrishnan, Renata Leite, Cathy L Melvin, Roger B Fillingim, Christie Frick, and Kathleen T Brady. 2016. Dental opioid prescribing and multiple opioid prescriptions among dental patients: administrative data from the south carolina prescription drug monitoring program. *The Journal of the American Dental Association*, 147(7):537–544.

Stephen E Nadeau, Jeffrey K Wu, and Richard A Lawhern. 2021. Opioids and chronic pain: an analytic review of the clinical evidence. *Frontiers in Pain Research*, page 44.

National Institute on Drug Abuse. 2023. Aberrant drug taking behaviors information sheet. https://nida.nih.gov/sites/default/files/AberrantDrugTakingBehaviors.pdf. Accessed: 2023-04-14.

Brice Ozenne, Fabien Subtil, and Delphine Maucort-Boulch. 2015. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859.

Steven D Passik and Kenneth L Kirsh. 2007. Commentary on jung and reidenberg's "physicians being deceived": aberrant drug-taking behaviors: what pain physicians can know (or should know). *Pain Medicine*, 8(5):442–444.

Leonard J Paulozzi, Edwin M Kilbourne, and Hema A Desai. 2011. Prescription drug monitoring programs and death rates from drug overdose. *Pain medicine*, 12(5):747–754.

Russell K Portenoy. 1996. Opioid therapy for chronic nonmalignant pain: a review of the critical issues. *Journal of pain and symptom management*, 11(4):203–217.

Kathryn Rough, Krista F Huybrechts, Sonia Hernandez-Diaz, Rishi J Desai, Elisabetta Patorno, and Brian T Bateman. 2019. Using prescription claims to detect aberrant behaviors with opioids: comparison and validation of 5 algorithms. *Pharmacoepidemiology and drug safety*, 28(1):62–69.

Abeed Sarker, Graciela Gonzalez-Hernandez, Yucheng Ruan, and Jeanmarie Perrone. 2019. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA network open*, 2(11):e1914672–e1914672.

Lisa A Schloff, Thomas S Rector, Raafat Seifeldin, and J David Haddox. 2004. Identifying controlled substance patterns of utilization requiring evaluation using administrative claims data. *The American journal of managed care*, 10:783–790.

Jade Singleton, Chengxi Li, Peter D Akpunonu, Erin L Abner, and Anna M Kucharska-Newton. 2023. Using natural language processing to identify opioid use disorder in electronic health record data. *International Journal of Medical Informatics*, 170:104963.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Mark D Sullivan, Mark J Edlund, Ming-Yu Fan, Andrea DeVries, Jennifer Brennan Braden, and Bradley C Martin. 2010. Risks for possible and probable opioid misuse among recipients of chronic opioid therapy in commercial and medicaid insurance plans: The troup study. *Pain*, 150(2):332–339.

Eric C Sun, Anjali Dixit, Keith Humphreys, Beth D Darnall, Laurence C Baker, and Sean Mackey. 2017. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis. *bmj*, 356.

The Office of the National Coordinator for Health Information Technology. 2020. Health care provider definition and cross-reference table. https://www.healthit.gov/sites/default/files/page2/2020-08/Health_Care_Provider_Definitions_v3.pdf. Accessed: 2023-9-24.

Katie Thomas and Charles Ornstein. 2017. Amid opioid crisis, insurers restrict pricey, less addictive painkillers. *New York Times*, pages 463–84.

CG Tudor. 2013. Memorandum: Medicare part d overutilization monitoring system. *Centers for Medicare and Medicaid Services*.

U.S. Department of Health and Human Service. 2021. Does a physician need a patient's written authorization to send a copy of the patient's medical record to a specialist or other health care provider who will treat the patient? https://www.hhs.gov/hipaa/for-professionals/faq/271/does-a-physician-need-written-authorization-to-send-medical-records-to-a-specialist/index.html. Accessed: 2023-9-24.

Xun Wang. 2022. Using natural language processing to detect opioid-related aberrant behaviors from electronic health records. In *APHA 2022 Annual Meeting and Expo*. APHA.

Matthijs J Warrens. 2015. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5(4):1.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344.

Lynn R Webster and Rebecca M Webster. 2005. Predicting aberrant behaviors in opioid-treated patients: preliminary validation of the opioid risk tool. *Pain medicine*, 6(6):432–442.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Audrey J Weiss, Kevin C Heslin, Carol Stocks, and Pamela L Owens. 2020. Hospital inpatient stays related to opioid use disorder and endocarditis, 2016: statistical brief# 256.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. 2023. Multi-label few-shot icd coding as autoregressive generation with prompt. In *AAAI 2023*.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot icd coding. In *Findings of Empirical Methods in Natural Language Processing (EMNLP)*.

Hamada M Zahera, Ibrahim A Elgendy, Rricha Jalota, Mohamed Ahmed Sherif, and E Voorhees. 2019. Fine-tuned bert model for multi-label tweets classification. In *TREC*, pages 1–7.

Vivienne J Zhu, Leslie A Lenert, Kelly S Barth, Kit N Simpson, Hong Li, Michael Kopscik, and Kathleen T Brady. 2022. Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy. *Health Informatics Journal*, 28(2):14604582221107808.

A Details on Data Construction

A.1 Details on Data Collection

Table 9: Opioids and their generic naming that used for filtering.

Medication Names	Generic Names
Ascomp with Codeine	aspirin/butalbital/caffeine/codeine
B & O Supprettes	belladonna/opium
Darvon Compound-65	aspirin/caffeine/propoxyphene
Lorcet	acetaminophen/hydrocodone
Maxidone	acetaminophen/hydrocodone
Fiorinal with Codeine III	aspirin/butalbital/caffeine/codeine
Magnacet	acetaminophen/oxycodone
Meprozine	meperidine/promethazine
Fiorinal with Codeine	aspirin/butalbital/caffeine/codeine
Fioricet with Codeine	acetaminophen/butalbital/caffeine/codeine
Lorcet Plus	acetaminophen/hydrocodone
Percocet 10 / 325	acetaminophen/oxycodone
Primlev	acetaminophen/oxycodone
Suboxone	buprenorphine/naloxone
Ibudone	hydrocodone/ibuprofen
Lorcet 10 / 650	acetaminophen/hydrocodone
Panlor DC	acetaminophen/caffeine/dihydrocodeine
Reprexain	hydrocodone/ibuprofen
Percocet	acetaminophen/oxycodone
Combunox	ibuprofen/oxycodone
Hydrocet	acetaminophen/hydrocodone
Roxicet	acetaminophen/oxycodone
Tylox	acetaminophen/oxycodone
Xolox	acetaminophen/oxycodone
Vicodin ES	acetaminophen/hydrocodone
Hycet	acetaminophen/hydrocodone
Talacen	acetaminophen/pentazocine
Vicodin HP	acetaminophen/hydrocodone
Vicoprofen	hydrocodone/ibuprofen
Percocet 7.5 / 325	acetaminophen/oxycodone
Lortab	acetaminophen/hydrocodone
Norco	acetaminophen/hydrocodone
Vicodin	acetaminophen/hydrocodone
Percocet 5 / 325	acetaminophen/oxycodone
Stagesic	acetaminophen/hydrocodone
Targiniq ER	naloxone/oxycodone
Xodol	acetaminophen/hydrocodone
Endocet	acetaminophen/oxycodone
Ultracet	acetaminophen/tramadol
Panlor SS	acetaminophen/caffeine/dihydrocodeine
Zubsolv	buprenorphine/naloxone
Xartemis XR	acetaminophen/oxycodone
Talwin Nx	naloxone/pentazocine
Tylenol with Codeine	acetaminophen/codeine
Anexsia	acetaminophen/hydrocodone
Darvocet-N 50	acetaminophen/propoxyphene
Liquicet	acetaminophen/hydrocodone
Darvocet-N 100	acetaminophen/propoxyphene
Trezix	acetaminophen/caffeine/dihydrocodeine
Percodan	aspirin/oxycodone
Darvocet A500	acetaminophen/propoxyphene
Percocet 2.5 / 325	acetaminophen/oxycodone
Balacet	acetaminophen/propoxyphene
Aceta w/ Codeine	acetaminophen/codeine
Zamicet	acetaminophen/hydrocodone
Embeda	morphine/naltrexone
Bunavail	buprenorphine/naloxone
Tylenol with Codeine #3	acetaminophen/codeine
Narvox	acetaminophen/oxycodone
Zydone	acetaminophen/hydrocodone
Tylenol with Codeine #4	acetaminophen/codeine

Continued on next page

Table 9: continued from previous page

Medication Names	Generic Names
Capital w/ Codeine	acetaminophen/codeine
Co-Gesic	acetaminophen/hydrocodone
Cocet Plus	acetaminophen/codeine
Codrix	acetaminophen/codeine
Dolacet	acetaminophen/hydrocodone
Dolagesic	acetaminophen/hydrocodone
Endodan	aspirin/oxycodone
Perloxx	acetaminophen/oxycodone
Phrenilin with Caffeine and Codeine	acetaminophen/butalbital/caffeine/codeine
Roxilox	acetaminophen/oxycodone
Synalgos-DC	aspirin/caffeine/dihydrocodeine
Theracodophen Low 90	acetaminophen/hydrocodone
Tramapap	acetaminophen/tramadol
Trycet	acetaminophen/propoxyphene
Verdrocet	acetaminophen/hydrocodone
Zolvit	acetaminophen/hydrocodone
Astramorph PF	morphine
Ionsys	fentanyl
Lazanda	fentanyl
Levo-Dromoran	levorphanol
Numorphan	oxymorphone
Onsolis	fentanyl
Oxyfast	oxycodone
Palladone	hydromorphone
Roxanol	morphine
Roxanol-T	morphine
Roxicodone Intensol	oxycodone
Meperitab	meperidine
Methadone Diskets	methadone
Actiq	fentanyl
Fentora	fentanyl
Subutex	buprenorphine
Demerol	meperidine
Dolophine	methadone
Roxicodone	oxycodone
Duragesic-25	fentanyl
Infumorph	morphine
Methadose	methadone
Ultram ODT	tramadol
Dilaudid	hydromorphone
Subsys	fentanyl
MSIR	morphine
OxyContin	oxycodone
Paregoric	opium
Duragesic-100	fentanyl
Abstral	fentanyl
Oxydose	oxycodone
Stadol	butorphanol
Duragesic	fentanyl
Duragesic-50	fentanyl
Buprenex	buprenorphine
Zohydro ER	hydrocodone
Duragesic-75	fentanyl
MS Contin	morphine
Kadian	morphine
Opana	oxymorphone
Opana ER	oxymorphone
Sublimaze	fentanyl
Exalgo	hydromorphone
Opium Deodorized	opium
Oxaydo	oxycodone
Avinza	morphine
Nucynta ER	tapentadol
Darvon-N	propoxyphene
OxyIR	oxycodone
Nubain	nalbuphine
Dilaudid-HP	hydromorphone

Continued on next page

Table 9: continued from previous page

Medication Names	Generic Names
Rybix ODT	tramadol
Ultram ER	tramadol
Butrans	buprenorphine
Darvon	propoxyphene
Oramorph SR	morphine
Nucynta	tapentadol
Ultram	tramadol
Duramorph	morphine
Ryzolt	tramadol
Talwin	pentazocine
Duragesic-12	fentanyl
Alfenta	alfentanil
ConZip	tramadol
Hysingla ER	hydrocodone
Belbuca	buprenorphine
Dazidox	oxycodone
DepoDur	morphine liposomal
ETH-Oxydose	oxycodone
Oxecta	oxycodone
Probuphine	buprenorphine
RMS	morphine
Sufenta	sufentanil
Ultiva	remifentanil

Continued on next page

Table 10: ICD 9 and ICD 10 diagnosis codes relevant to OUD. Note that, all of these codes defined by Weiss et al. (2020).

ICD code	ICD Description
	ICD 9 diagnosis codes
304	Opioid type dependence, unspecified
304.01	Opioid type dependence, continuous
304.02	Opioid type dependence, episodic
304.03	Opioid type dependence, in remission
304.7	Combinations of opioid type drug with any other drug dependence, unspecified
304.71	Combinations of opioid type drug with any other drug dependence, continuous
304.72	Combinations of opioid type drug with any other drug dependence, episodic
304.73	Combinations of opioid type drug with any other drug dependence, in remission
305.5	Opioid abuse, unspecified
305.51	Opioid abuse, continuous
305.52	Opioid abuse, episodic
305.53	Opioid abuse, in remission
965	Poisoning by opium (alkaloids), unspecified
965.01	Poisoning by heroin
965.02	Poisoning by methadone
965.09	Poisoning by other opiates and related narcotics
970.1	Poisoning by opiate antagonists
E850.0	Accidental poisoning by heroin
E850.1	Accidental poisoning by methadone
E850.2	Accidental poisoning by other opiates and related narcotics
E935.0	Heroin causing adverse effects in therapeutic use
E935.1	Methadone causing adverse effects in therapeutic use
E935.2	Other opiates and related narcotics causing adverse effects in therapeutic use
E940.1	Adverse effects of opiate antagonists
	ICD 10 diagnosis codes
Opioid abuse/dependence	
F11.10	Opioid abuse, uncomplicated
F11.120	Opioid abuse with intoxication, uncomplicated
F11.121	Opioid abuse with intoxication, delirium
F11.122	Opioid abuse with intoxication, with perceptual disturbance
F11.129	Opioid abuse with intoxication, unspecified
F11.14	Opioid abuse with opioid-induced mood disorder
F11.150	Opioid abuse with opioid-induced psychotic disorder, with delusions
F11.151	Opioid abuse with opioid-induced psychotic disorder, with hallucinations
F11.159	Opioid abuse with opioid-induced psychotic disorder, unspecified
F11.181	Opioid abuse with opioid-induced sexual dysfunction
F11.182	Opioid abuse with opioid-induced sleep disorder
F11.188	Opioid abuse with other opioid-induced disorder
F11.19	Opioid abuse with unspecified opioid-induced disorder
F11.20	Opioid dependence, uncomplicated
F11.21	Opioid dependence, in remission
F11.220	Opioid dependence with intoxication, uncomplicated
F11.221	Opioid dependence with intoxication, delirium
F11.222	Opioid dependence with intoxication, with perceptual disturbance
F11.229	Opioid dependence with intoxication, unspecified
F11.23	Opioid dependence with withdrawal
F11.24	Opioid dependence with opioid-induced mood disorder
F11.250	Opioid dependence with opioid-induced psychotic disorder, with delusions
F11.251	Opioid dependence with opioid-induced psychotic disorder, with hallucinations
F11.259	Opioid dependence with opioid-induced psychotic disorder, unspecified
F11.281	Opioid dependence with opioid-induced sexual dysfunction
F11.282	Opioid dependence with opioid-induced sleep disorder
F11.288	Opioid dependence with other opioid-induced disorder
F11.29	Opioid dependence with unspecified opioid-induced disorder
Opioid use	
F11.90	Opioid use, unspecified, uncomplicated
F11.920	Opioid use, unspecified with intoxication, uncomplicated
F11.921	Opioid use, unspecified with intoxication delirium
F11.922	Opioid use, unspecified with intoxication, with perceptual disturbance
F11.929	Opioid use, unspecified with intoxication, unspecified
F11.93	Opioid use, unspecified, with withdrawal
F11.94	Opioid use, unspecified, with opioid-induced mood disorder
F11.950	Opioid use, unspecified with opioid-induced psychotic disorder, with delusions

Continued on next page

Table 10: continued from previous page

Diagnosis code	Description
F11.951	Opioid use, unspecified with opioid-induced psychotic disorder, with hallucinations
F11.959	Opioid use, unspecified with opioid-induced psychotic disorder, unspecified
F11.981	Opioid use, unspecified with opioid-induced sexual dysfunction
F11.982	Opioid use, unspecified with opioid-induced sleep disorder
F11.988	Opioid use, unspecified with other opioid-induced disorder
F11.99	Opioid use, unspecified, with unspecified opioid-induced disorder
Poisoning	
T40.0X1A	Poisoning by opium, accidental (unintentional), initial encounter
T40.0X1D	Poisoning by opium, accidental (unintentional), subsequent encounter
T40.0X2A	Poisoning by opium, intentional self-harm, initial encounter
T40.0X2D	Poisoning by opium, intentional self-harm, subsequent encounter
T40.0X3A	Poisoning by opium, assault, initial encounter
T40.0X3D	Poisoning by opium, assault, subsequent encounter
T40.0X4A	Poisoning by opium, undetermined, initial encounter
T40.0X4D	Poisoning by opium, undetermined, subsequent encounter
T40.1X1A	Poisoning by heroin, accidental (unintentional), initial encounter
T40.1X1D	Poisoning by heroin, accidental (unintentional), subsequent encounter
T40.1X2A	Poisoning by heroin, intentional self-harm, initial encounter
T40.1X2D	Poisoning by heroin, intentional self-harm, subsequent encounter
T40.1X3A	Poisoning by heroin, assault, initial encounter
T40.1X3D	Poisoning by heroin, assault, subsequent encounter
T40.1X4A	Poisoning by heroin, undetermined, initial encounter
T40.1X4D	Poisoning by heroin, undetermined, subsequent encounter
T40.2X1A	Poisoning by other opioids, accidental (unintentional), initial encounter
T40.2X1D	Poisoning by other opioids, accidental (unintentional), subsequent encounter
T40.2X2A	Poisoning by other opioids, intentional self-harm, initial encounter
T40.2X2D	Poisoning by other opioids, intentional self-harm, subsequent encounter
T40.2X3A	Poisoning by other opioids, assault, initial encounter
T40.2X3D	Poisoning by other opioids, assault, subsequent encounter
T40.2X4A	Poisoning by other opioids, undetermined, initial encounter
T40.2X4D	Poisoning by other opioids, undetermined, subsequent encounter
T40.3X1A	Poisoning by methadone, accidental (unintentional), initial encounter
T40.3X1D	Poisoning by methadone, accidental (unintentional), subsequent encounter
T40.3X2A	Poisoning by methadone, intentional self-harm, initial encounter
T40.3X2D	Poisoning by methadone, intentional self-harm, subsequent encounter
T40.3X3A	Poisoning by methadone, assault, initial encounter
T40.3X3D	Poisoning by methadone, assault, subsequent encounter
T40.3X4A	Poisoning by methadone, undetermined, initial encounter
T40.3X4D	Poisoning by methadone, undetermined, subsequent encounter
T40.4X1A	Poisoning by synthetic narcotics, accidental (unintentional), initial encounter
T40.4X1D	Poisoning by synthetic narcotics, accidental (unintentional), subsequent encounter
T40.4X2A	Poisoning by other synthetic narcotics, intentional self-harm, initial encounter
T40.4X2D	Poisoning by other synthetic narcotics, intentional self-harm, subsequent encounter
T40.4X3A	Poisoning by other synthetic narcotics, assault, initial encounter
T40.4X3D	Poisoning by other synthetic narcotics, assault, subsequent encounter
T40.4X4A	Poisoning by synthetic narcotics, undetermined, initial encounter
T40.4X4D	Poisoning by synthetic narcotics, undetermined, subsequent encounter
T40.601A	Poisoning by unspecified narcotics, accidental (unintentional), initial encounter
T40.601D	Poisoning by unspecified narcotics, accidental (unintentional), subsequent encounter
T40.602A	Poisoning by unspecified narcotics, intentional self-harm, initial encounter
T40.602D	Poisoning by unspecified narcotics, intentional self-harm, subsequent encounter
T40.603A	Poisoning by unspecified narcotics, assault, initial encounter
T40.603D	Poisoning by unspecified narcotics, assault, subsequent encounter
T40.604A	Poisoning by unspecified narcotics, undetermined, initial encounter
T40.604D	Poisoning by unspecified narcotics, undetermined, subsequent encounter
T40.691A	Poisoning by other narcotics, accidental (unintentional), initial encounter
T40.691D	Poisoning by other narcotics, accidental (unintentional), subsequent encounter
T40.692A	Poisoning by other narcotics, intentional self-harm, initial encounter
T40.692D	Poisoning by other narcotics, intentional self-harm, subsequent encounter
T40.693A	Poisoning by other narcotics, assault, initial encounter
T40.693D	Poisoning by other narcotics, assault, subsequent encounter
T40.694A	Poisoning by other narcotics, undetermined, initial encounter
T40.694D	Poisoning by other narcotics, undetermined, subsequent encounter
Adverse effects	
T40.0X5A	Adverse effect of opium, initial encounter
T40.0X5D	Adverse effect of opium, subsequent encounter
T40.2X5A	Adverse effect of other opioids, initial encounter

Continued on next page

Table 10: continued from previous page

Diagnosis code	Description
T40.2X5D	Adverse effect of other opioids, subsequent encounter
T40.3X5A	Adverse effect of methadone, initial encounter
T40.3X5D	Adverse effect of methadone, subsequent encounter
T40.4X5A	Adverse effect of synthetic narcotics, initial encounter
T40.4X5D	Adverse effect of synthetic narcotic, subsequent encounter
T40.605A	Adverse effect of unspecified narcotics, initial encounter
T40.605D	Adverse effect of unspecified narcotics, subsequent encounter
T40.695A	Adverse effect of other narcotics, initial encounter
T40.695D	Adverse effect of other narcotics, subsequent encounter
Long-term use of opiates	
Z79.891	Long-term (current) use of opiate analgesic

Continued on next page

A.2 Detailed Descriptions on the Categories

Confirmed aberrant behavior (CAB): This class refers to behavior that more likely lead to a catastrophic adverse events. It is defined as evidence confirming loss of control of opioid use, specifically aberrant usage of opioid medications, including: 1) Aberrant use of opioids, such as administration/consumption in a way other than described or self-escalating doses. 2) Evidence suggesting or proving that patient has been selling or giving away opioids to others, including family members. 3) Use of opioids for a different indication other than the indication intended by the prescriber. 4) Phrases suggesting current use of illicit or illicitly obtained substances or misuse of legal substances (e.g. alcohol) other than prescription opioid medications.

Suggested aberrant behavior (SAB): This class refers to behavior implying patient distress related to their opioid treatment. SAB includes three kinds of behavior that suggest potential misuse of opioid. 1) Patient attempt to get extra opioid medicine like requesting for early refill, asking for increasing dosage or reporting missing/stolen opioid medication. 2) Patient emotions toward opioid like request of a certain opioid medication use/change/increase. 3) Physician concerns.

Opioids: This class refers to the mention or listing of the name(s) of the opioid medication(s) that the patient is currently prescribed or has just been newly prescribed.

Indication: This class indicates that patients are using opioid under instructions, such as using opioid for pain, for treatment of opioid use disorder, etc.

Opioid dependence: It refers to patients have the condition of being dependent on opioids, have chronic opioid use, or is undergoing opioid titration.

Benzodiazepines: This class refers co-prescribed benzodiazepines (a risk factor for accidental opioid overdose (Sun et al., 2017)). In this case, the patient is simply being co-prescribed benzodiazepines (with no noted evidence for abuse).

Medication Change: This class indicates that the physician makes changes to the patient's opioid regimen during this current encounter or the patient's opioid regimen has been changed since the patient's last encounter with the provider writing the note.

Central Nervous System Related: This is defined as CNS-related terms or terms suggesting altered sensorium, including cognitive impairment, sedation, lightheadedness, intoxication and general term suggesting altered sensorium (e.g. "altered mental status").

Social Determinants of Health: This class refers to the factors in the surroundings which impact their well-being. Our dataset captured following attributes:

- Marital status (single, married ...)
- Cohabitation status (live alone, lives with others ...)
- Educational level (graduate degree, college degree, high-school diploma ...)
- Socioeconomic status (retired, disabled, pension, working ...)
- Homelessness (past, present ...)

B Details on the Data Augmentation with a Large Language Model

Categories	BioClinicalBERT		T5 Paraphrasing	
	AUPRC	F1	AUPRC	F1
Confirmed Aberrant Behaviors	88.11±7.94	72.25±8.50	**91.03±6.29**	**85.29±6.26**
Suggested Aberrant Behaviors	46.88±11.23	49.70±10.6	**62.66±14.71**	**53.38±13.17**
Opioids	**99.52±0.20**	97.65±0.53	99.34±0.39	**97.94±0.35**
Indication	**97.77±0.78**	93.37±1.60	96.68±1.17	**95.12±1.62**
Diagnosed Opioid Dependency	84.20±7.58	70.23±15.04	**92.84±5.42**	**86.91±8.54**
Benzodiazepines	**96.68±1.11**	**97.31±0.51**	95.92±2.23	96.65±0.94
Medication Change	75.25±4.35	67.20±4.79	**77.14±5.04**	**72.48±3.91**
Central nervous system related	98.18±0.67	90.23±1.49	**98.83±0.67**	**94.91±0.93**
Social Determinants of Health	95.68±1.76	91.45±2.63	**96.28±2.76**	**95.89±1.77**
Macro Average	86.92±17.04	81.04±16.78	**90.08±12.28**	**86.51±14.84**

Table 11: Experimental results of the data augmentation with a LLM's paraphrasing.

Experimental results in Table 11 showed that the data augmentation helps to enhance the performance of aberrant behavior detection at BioClinicalBERT + Prompt-based training environment. Especially the performance of the uncommon classes, such as diagnosed opioid dependence, suggested aberrant behaviors, diagnosed opioid dependency, increased significantly. However, if there is already enough data and performance is high (Opioids, Indication, Benzodiazepines, Central nervous systerm related, Social determinant of health), there is a marginal difference in performance. In addition, due to the various linguistic patterns of suggested aberrant behaviors, there is still room for performance improvement by paraphrasing alone although the performance enhanced significantly.