Computer Methods and Programs in Biomedicine 236 (2023) 107573

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Machine learning for predicting opioid use disorder from healthcare data: A systematic review☆

Christian Garbin*, Nicholas Marques, Oge Marques

Department of Electrical Engineering and Computer Science, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA

ARTICLE INFO

Article history:
Received 8 February 2023
Revised 16 April 2023
Accepted 26 April 2023

Keywords:
Opioid use disorder (OUD)
Systematic review
Machine learning
Artificial intelligence

ABSTRACT

Introduction: The US opioid epidemic has been one of the leading causes of injury-related deaths according to the CDC Injury Center. The increasing availability of data and tools for machine learning (ML) resulted in more researchers creating datasets and models to help analyze and mitigate the crisis. This review investigates peer-reviewed journal papers that applied ML models to predict opioid use disorder (OUD). The review is split into two parts. The first part summarizes the current research in OUD prediction with ML. The second part evaluates how ML techniques and processes were used to achieve these results and suggests improvements to refine further attempts to use ML for OUD prediction.

Methods: The review includes peer-reviewed journal papers published on or after 2012 that use healthcare data to predict OUD. We searched Google Scholar, Semantic Scholar, PubMed, IEEE Xplore, and Science.gov in September of 2022. Data extracted includes the study's goal, dataset used, cohort selected, types of ML models created, model evaluation metrics, and the details of the ML tools and techniques used to create the models.

Results: The review analyzed 16 papers. Three papers created their dataset, five used a publicly available dataset, and the remaining eight used a private dataset. Cohort size ranged from the low hundreds to over half a million. Six papers used one type of ML model, and the remaining ten used up to five different ML models. The reported ROC AUC was higher than 0.8 for all but one of the papers. Five papers used only non-interpretable models, and the other 11 used interpretable models exclusively or in combination with non-interpretable ones. The interpretable models were the highest or second-highest ROC AUC values. Most papers did not sufficiently describe the ML techniques and tools used to produce their results. Only three papers published their source code.

Conclusions: We found that while there are indications that ML methods applied to OUD prediction may be valuable, the lack of details and transparency in creating the ML models limits their usefulness. We end the review with recommendations to improve studies on this critical healthcare subject.

Abbreviations

OUD	Opioid Use Disorder
ML	Machine Learning
PRISMA	Preferred reporting Items for Systematic Reviews and Meta-Analyses
ROC AUC	Receiver Operating Characteristic Area Under the Curve
PPV	Positive Predictive Value
TPR	True Positive Rate
AUPRC	Area Under the Precision-Recall Curve

1. Introduction

Since the early 2000s, opioid use disorder and overdose rates have skyrocketed in the United States [1]. Whether opioid use begins through prescription or illicit routes, opioid use disorder and overdose rates led the United States to consider it an epidemic and declare a public health emergency in 2017 [2]. In recent years, as the healthcare community searches for more effective ways to mitigate the opioid crisis, there have been rapid advancements in machine learning. The availability of more data and better machine learning (ML) frameworks has led to the development of ML models that use healthcare data to deal with different facets of the opioid crisis. We examined peer-reviewed papers that apply ML methods to one aspect of the opioid crisis, the prediction of opioid use disorder (OUD).

☆ Review Board or equivalent approval or exemption: not required for this work.

* Corresponding author.
E-mail address: cgarbin@fau.edu (C. Garbin).

https://doi.org/10.1016/j.cmpb.2023.107573
0169-2607/

With the proliferation of ML frameworks and tools, it is now easy to create models with a few lines of code that perform well in a restricted setting. However, ML models created without following the best practices in data science and machine learning do not advance the field because their results are unreliable and not reproducible. Therefore, in this systematic review, we analyzed not only the results from the models but also the ML methods used to preprocess the datasets, create, and evaluate the models. We end the review with recommendations for future studies that use machine learning.

The closest related work in this area are [3], which investigates the use of ML in addiction in general (not only OUD), and [4], which investigates the use of ML models in different aspects of opioid usage, including risk prediction and pain management. This systematic review differs from the related works by analyzing the technical aspects of creating the ML models in addition to their results.

2. Methods

We conducted a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations. Two questions guided the review: (1) What machine learning methods are being applied to OUD prediction, and what are the results of these models? (2) What machine learning practices do the papers apply to process the dataset, train, and evaluate the models to report their results?

2.1. Inclusion criteria

We included papers that met the following criteria: published in a peer-reviewed journal, used one or more machine learning models (ML) to predict opioid use disorder (OUD), must be exclusively about OUD and not general drug abuse, must be specifically about use disorder and not related issues such as prolonged use, and must use healthcare data. We included papers published in or after 2012, when the seminal AlexNet brought deep neural networks into the mainstream and machine learning libraries started to become more accessible for general use [5].

2.2. Exclusion criteria

We excluded papers that combined opioids with other drugs (e.g. papers that mixed opioid and alcohol or marijuana use were not eligible), papers that did not use healthcare data (e.g. data mined from social media or similar sources), papers that developed ML models for survival analysis (as opposed to prediction), and papers that were about the legitimate use of opioids, not abuse.

2.3. Search strategy and study selection

We searched the following databases in September of 2022: Semantic Scholar, Google Scholar, Science.gov, IEEE Xplore, and PubMed. After initial queries, we removed studies based on title, abstract and keyword, type, and full-text review. Two authors reviewed each item. Appendix A lists the exact keywords and filters used in the queries. Appendix B describes the review process in detail.

Fig. 1 shows the flowchart of the paper selection process. After duplicate removal, we reviewed the title of 1320 papers and selected 158 for an abstract and keyword review, from which we selected 52 for a type filter (peer-reviewed or not), which left 32 papers for a full-text review. The final selection resulted in the 16 papers analyzed in this review.

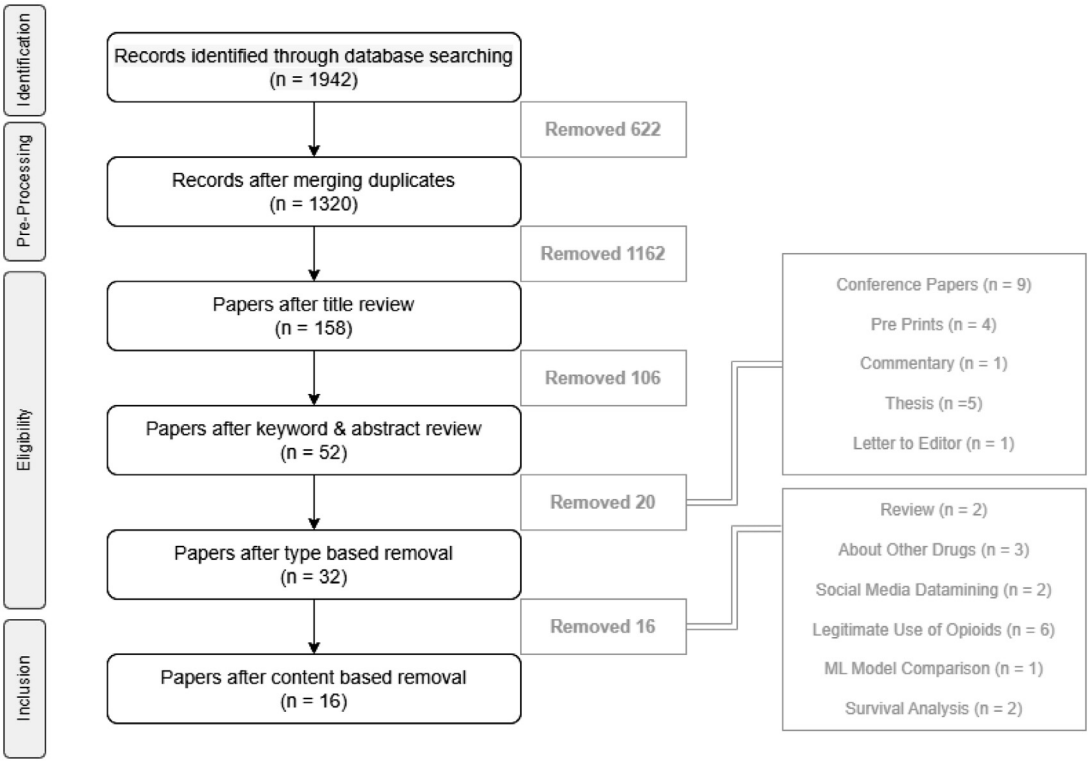

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart. The chart shows the number of papers removed in each filtering step. Initial preprocessing was done with Zoteros's duplicate removal function, followed by a manual check for duplicates. Eligibility was determined in multiple rounds, as shown in the figure. Two authors participated in all steps, deciding by consensus. The review process is described in detail in Appendix B.

Table 1
Papers reviewed, journal and year of publication, and stated objective.

Paper	Title	Journal/year	Stated objective
[9]	Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence	Drug and Alcohol Dependence 2016	Identify substance-specific behavioral markers for heroin and amphetamine dependence with "demographic, personality, psychiatric, and neurocognitive measures of impulsivity and related constructs from individuals with lifetime mono-dependence on heroin or amphetamine, lifetime polysubstance dependence, and no history of dependence".
[10]	The Opioid Abuse Risk Screener predicts aberrant same-day urine drug tests and 1-year controlled substance database checks: a brief report	Health Psychology Open 2017	Evaluate the predictive validity of the OARS (Opioid abuse risk screener) and "test the feasibility of using [an ML] algorithm to evaluate psychometric properties of the OARS in a small-to-moderate sample size, similar to traditional psychiatric research populations."
[11]	Convergence of case-specific epigenetic alterations identify a confluence of genetic vulnerabilities tied to opioid overdose	Molecular Psychiatry 2022	Use an ML model to find "shortcut discovery of genes involved in the neurobiology of OUD."
[12]	Identifying risk of opioid use disorder for patients taking opioid medications with deep learning	Journal of the American Medical Informatics Association 2021	Develop and evaluate models to "predict OUD for patients on opioid medications using electronic health records and deep learning methods."
[13]	Predicting opioid dependence from electronic health records with machine learning	BioData Mining 2019	Train an ML model to "classify patients by likelihood of having a diagnosis of substance dependence using EHR data from patients diagnosed with substance dependence".
[14]	Predicting opioid use disorder and associated risk factors in a Medicaid managed care population	American Journal Of Managed Care 2021	Develop and validate "a predictive model of OUD and to predict future OUD diagnosis"
[15]	Using machine learning to predict opioid misuse among U.S. adolescents	Preventive Medicine 2020	Evaluate "opioid misuse prediction performance of three different ML techniques" and "[c]ompare such performance with the performance of the logistic regression in a nationally representative sample of U.S. adolescents"
[6]	A machine learning framework to predict the risk of opioid use disorder	Machine Learning with Applications 2021	Identify "potential risk factors of opioid use disorder from a large-scale healthcare claims data"
[8]	Using machine learning to predict risk of incident opioid use disorder among fee-for-service Medicare beneficiaries: a prognostic study	PLOS ONE 2020	Develop and validate an ML model to "predict incident OUD among Medicare beneficiaries having at least one opioid prescription."
[16]	Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing	Frontiers in Public Health 2022	Develop "an annotation schema to deeply characterize OUD, and to automate the schema using machine learning and deep learning-based approaches" and "present the results of two supervised classification approaches [based on the schema]."
[17]	The detection of opioid misuse and heroin use from paramedic response documentation: machine learning for improved surveillance	Journal of Medical Internet Research 2020	Develop and test "a natural language processing method that would improve identification of potential OM [opioid misuse] from paramedic documentation."
[18]	Development of a machine learning algorithm for early detection of opioid use disorder	Pharmacology Research & Perspectives 2020	Create a prediction model and algorithm to "identify patients at high risk for OUD before OUD has been fully developed and diagnosed, in order to be able to offer them early prevention and interventions."
[19]	Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients	BMC Medical Informatics and Decision Making 2020	Compare the performance "of multiple text classification approaches, including both PHI-laden and PHI-free [PHI=protected health information], for an opioid misuse".
[20]	Clinical prediction of extra-medical use of prescription pain relievers from a representative United States sample	Preventive Medicine 2021	Identify "patients at high risk of EMPPR [extra-medical use of prescription pain relievers], who need increased monitoring of dispensed opioids, such as with 'opioid use contracts'."
[7]	Predictive modeling of susceptibility to substance abuse, mortality and drug-drug interactions in opioid patients	Frontiers in Artificial Intelligence 2021	Present "a collection of predictive models to identify patients at risk of opioid abuse and mortality by using their prescription histories."
[21]	Understanding opioid use disorder (OUD) using tree-based classifiers	Drug and Alcohol Dependence 2020	Develop and compare ML approaches to " predict individuals that are at risk for OUD and to understand how interactions between various demographic, socioeconomic, physical, and psychological predictors increase this risk."

EHR: electronic health records, ML: machine learning, OUD: opioid use disorder.

3. Results

3.1. Characteristics of included studies

Table 1 lists the selected papers, their title, the journal where they were published, the publication year, and their stated objective. Two papers [6,7] were published in computer science journals, one paper was published in a multidisciplinary journal [8], and the other 13 papers were published in medical, healthcare, or biology journals. The earliest paper is from 2016 [9] and 13 papers are from 2020 or later. To check if our selection process favored recent papers, we analyzed the dates of the non-duplicate items identified in the initial search. We concluded that papers in this area have primarily been published in recent years (Appendix D).

Table 2 shows each paper's dataset, population, years covered, cohort selection, and cohort size (number of people and number of health records, when available). Three papers created a dataset as part of their work [9–11], five studies used a publicly-available dataset (NSDUH or MIMIC-III) [7,15,16,20,21]. The remaining eight papers used a private dataset from a medical or healthcare organization. Through a combination of a small initial dataset and cohort selection, five papers used 1000 or fewer individuals for the ML train/test process [9–11,16,19], one paper did not specify the final size [21], and the remaining ten papers used 1000 more individuals. Seven papers [9–11,14,15,17,21] used data from the past ten years (2012 or newer), and the other nine papers used data older than ten years or a mixture of data older than years and data from the past ten years. Of the eight papers that mention age as one of the cohort selection criteria (inclusion or exclusion), two selected adolescents [15,20], and six selected adults (18 years or older) [6,8,9,12–14].

3.2. Main findings

3.2.1. Machine learning models

Table 3 shows that we can split the use of ML models as follows:

- *Number of models*: Six papers used only one type of ML model [9–11,13,18,20]. The other ten papers used more than one ML model.
- *Types of models*: Two papers used interpretable ML models (logistic regression in both cases) exclusively [9,20]. Nine papers used interpretable ML models (logistic regression or decision tree) and non-interpretable models (multiple types) [6–8,12,14–16,19,21]. The remaining five papers used non-interpretable ML models (various types) exclusively.

3.2.2. Evaluation metrics

Table 3 shows that most papers used ROC AUC (receiver operating characteristic area under the curve) as an evaluation metric (all but one paper [16] reported it), followed by precision (PPV, positive predictive value) and recall (TPR, true positive rate). Likely due to publication bias [22,23], papers reported high ROC AUC values, supporting their claims. The sole paper with a low ROC AUC [10] noted that it could be related to the small number of patients in the dataset. Excluding that paper, the minimum reported ROC AUC is 0.811 [15], and the maximum is 0.99 [7].

Of the papers that used more than one ML model, the interpretable model is either the best ROC AUC [15,17] or the second-best one [6–8,12,14,19,21].

3.2.3. Machine learning methods and reproducibility of the experiments

Over the years, the machine learning community has developed methods to improve model training, evaluation, and deployment. Table 4 summarizes how the papers employed or failed to employ these methods.

- *Class imbalance*: The incidence of OUD is small in the general population [24]. In ML terminology, it results in a dataset with "class imbalance," where the number of positive class (OUD) samples is small compared to the negative (non-OUD) class. Several techniques exist to train a model with class imbalance [25]. Eleven papers did not describe which technique they used or if they used any technique [8–12,14–18,20]. The remaining five papers [6,7,13,19,21] described what they used with varying degrees of detail.
- *Train/test set split*: ML models must be evaluated on unseen data, i.e. data they have not been trained on. In ML terminology, this is the "test set." The test set must be set aside early in the process before a model is trained, then used later to evaluate the model. The test set must match the characteristics of the population where the model will be later deployed [26]. Three papers did not split the dataset or did not describe if they did so [10,13,21]. Five papers split the dataset but did not describe the criteria used for the split [7,9,11,17,19]. Eight papers split the dataset and explained the criteria [6,8,12,14–16,18,20].
- *Reproducibility*: Reproducing an ML experiment requires a detailed description of the tools and parameters used for the original research [27,28]. First, all the hyperparameters used to train the model must be described. Then the exact version of the programming language compiler or interpreter and each library must be listed. The post-processed dataset (after data clean-up and missing data imputation) and the tools to create it must be available whenever possible (respecting privacy and ethical considerations). Finally, the code to train and test the model must be published [29]. None of the papers met all of the reproducibility requirements. Five papers described the training hyperparameters in detail [11,15,16,18,20], and four others described them partially [7,13,17,19]. None of the papers described the version of tools and libraries to the level needed to reproduce the results. Three papers made the code available outright [11,12,17]. One paper made the code available upon request [7]. The remaining 12 papers did not make the code available.

4. Discussion

Reliably predicting opioid use disorder (OUD) from healthcare data has immediate and measurable benefits for individuals and society. Machine learning models built on healthcare data are a promising part of the solution because they can be built on data already collected in the industry for other reasons, saving one of the most expensive parts of the ML process, procuring a dataset.

4.1. Challenges and recommendations

In this section, we discuss challenges to developing ML models of OUD prediction and list recommendations to improve future research on this topic. As reported in the related works [3] and [4], we also found out that most studies used supervised learning. Here we expand on that work by analyzing the technical aspects of the creation and evaluation of the ML models.

4.1.1. Appropriate model metrics

All papers reported performance with precision and recall, either the numbers, the number and the curve (ROC AUC), or both. However, only six papers reported AUPRC (area under the precision/recall curve), a metric more useful for imbalanced datasets with more negative class samples than the positive class [30]. We recommend reporting AUPRC in future research, given the low incidence of OUD in the general population [24].

Table 2
Dataset, the population of the dataset, years covered in the dataset, criteria to select cohort from the dataset, the size of the study in the number of people and medical records (when available) after cohort selection.

Paper	Dataset	Population	Years covered	Cohort selection	Cohort size (people/records)
[9]	Created in the study	222 volunteers enrolled in a larger study on impulsivity among drug users.	2016	Between 18 and 50 years, IQ > 75, 8th grade or higher education, no history of neurological illnesses, HIV negative, negative urine test for cannabis, amphetamine, methamphetamine, and opioids.	222/Not specified
[10]	Created in the Study	612 patients who completed an opioid abuse risk screener (OARS) as part of routine clinical practice.	2017	Urine drug test and OARS done on the same day, and controlled substance database (CSDB) within one year of the OARS.	532/Not specified
[11]	Created in the study	"[O]pportunistic sample of opioid-related deaths and unaffected controls that came to autopsy."	2022	Samples could not have a history of psychiatric disorders, debilitating chronic pain, or death by suicide.	102 (51 cases and 51 controls)/Not applicable
[12]	Cerner's Health Facts Database	"[P]atients who have been prescribed with medications containing active opioid ingredients".	2008-2017	Patients with at least one prescription of opioid medication, between 18 and 66 when first exposed to opioid medication, not being treated for cancer.	111,456 positive (OUD) and 5072,110 negative patients/Not specified
[13]	Mount Sinai Medical Center (MSMC) EHR	Patients with records in the MSMC EHR	2000-2015	Patients were excluded if they were diagnosed with substance dependence before age 20, with at least 17 recorded lab tests and vital signs. Outliers and typos were also removed.	7797 cases and 191,476 control patients/Not specified
[14]	Medicaid enrollment, medical, pharmacy, and care management administrative data from a private Medicaid managed care organization (AmeriHealth Caritas)	Records from DC, FL, LA, MI, PA, and SC.	2017-2019	Adults continuously enrolled in Medicaid in the covered years.	2017 (n = 320,040) 2018 (n = 374,809) 2019 (n = 589,423)/Not specified
[15]	NSDUH	The NSDUH population	2015-2017	Adolescents between 12 and 17 years.	41,579/Not applicable
[6]	Massachusetts All Payer Claim Datasets (MA APCD), a commercial insurance claims dataset	Patients from the MA APCD	2011 - 2013	Included patients continuously insured during the study time. Excluded age below 18 and records missing gender, both a small number of records	~600,000/"The pharmacy claims file contains data for approximately 470 million prescriptions and the medical claims file has approximately 1.63 billion claims"
[8]	Unspecified dataset from the Centers for Medicare and Medicaid Services (CMS) database from https://resdac.org/.	A 5% random sample of Medicare beneficiaries.	2011 - 2016	Included "fee-for-service adult beneficiaries aged >= 18 years who were US residents and received >= 1 non-parenteral and non-cough/cold opioid prescriptions." Excluded malignant cancer diagnosis, OUD diagnosis before initiating opioids, other substance use disorder.	361,527/Not specified
[16]	MIMIC-III	The MIMIC-III population	2001 - 2012	"[P]atients who had an International Classification of Diseases, version 9 (ICD-9) code related to OUD".	762/Not specified
[17]	Denver Health paramedic trip reports	The entire dataset	2017–8 - 2018–4	Records that included "keywords naloxone, heroin, and both combined"	Not specified/1298
[18]	"[A] commercial claims database of a large American health maintenance organization of over 20 million patients..."	Ten million insurance claims sampled from the dataset	2006 - 2018	Included "patients who purchased at least one medication from the opioid class for example after trauma or medical procedures, excluding codeine". Excluded "patients diagnosed with cancer or assigned palliative care [or] missing data from the 11 defining problems."	130,451/550,000
[19]	Chicago's Loyola University Medical Center (LUMC) EHR system	161,520 adult (age 18 or older) inpatient encounters	2007 - 2017	Random 1000 patients, oversampled for opioid-related hospitalizations or positive for urine opioid drug test.	1000/63,301
[20]	NSDUH	The NSDUH population	2004 - 2018	Adolescents between 12 and 17 years, excluding records with missing data.	234,593/Not applicable
[7]	MIMIC-III	The MIMIC-III population	2001 - 2012	Patients who were prescribed opioids, opiates, or naloxone.	29,959/Not specified
[21]	NSDUH	The NSDUH population	2016	Unclear ("A data set was curated from these survey responses.")	Unclear/Not applicable

EHR: electronic health record.
MIMIC-III: (Medical Information Mart for Intensive Care) is a dataset from the Beth Israel Deaconess Medical Center, Boston, Massachusetts, with data from 46,530 patients.
NSDUH: (National Survey on Drug Use and Health) is a yearly survey covering all 50 states and the District of Columbia in the United States, with approximately 70,000 participants aged 12 years and older.

Table 3
Machine learning models, metrics used to evaluate them, and best model (N/A if used only one model). Machine learning models are ordered from simpler (more interpretable) to more complex (less interpretable): RG: regression (logistic or Cox regression, with or without regularization), DT: decision tree, SVM: support vector machine, RF: random forest and random survival forest, GB: gradient boosting, KNN: k-nearest neighbor, NN: neural network, including MLPs (multilayer perceptron, dense networks), DNN (deep neural networks), CNN (convolutional neural networks), RNN (recurrent neural networks), and transformer-based networks. Evaluations metrics: TNR: true negative rate (specificity), TPR: true positive rate (recall, sensitivity), PPV: positive predictive value (precision), NPV: negative predictive value, ROC AUC: receiver operating characteristic area under the curve, AUPRC: area under precision-recall curve. Appendix C lists the models and metrics in detail, for example, what types of regressions (LASSO, Ridge, and others) and what type of neural networks (deep neural networks, convolutional neural networks, and others).

Paper	Machine learning models							Evaluation metrics						Best and second best model by ROC AUC
	RG	DT	SVM	RF	GB	KNN	NN	TNR	TPR	PPV	NPV	ROC AUC	AUPRC	
[9]	X											X		RG (0.870) N/A
[10]			X									X		SVM (0.626) N/A
[11]					X					X		X		GB (0.972) N/A
[12]	X	X		X			X		X	X		X		NN (0.9224) DT (0.8823)
[13]				X				X	X	X	X	X	X	RF (0.863) N/A
[14]	X	X	X	X			X			X		X		NN (0.918) RG (0.915)
[15]	X			X	X		X					X	X	RG (0.815) GB (0.811)
[6]	X	X		X	X				X	X		X		RF (0.97) DT (0.956)
[8]	X			X	X		X	X	X	X	X	X	X	GB (0.882) RG (0.880)
[16]	X						X		X	X				N/A
[17]			X	X		X		X	X	X	X	X		RG (0.94) RF (0.91)
[18]					X			X	X	X	X	X	X	GB (0.959) N/A
[19]	X						X	X	X	X	X	X		NN (0.94) RG (0.91)
[20]	X											X	X	RG (0.819) N/A
[7]	X				X				X	X	X	X	X	GB (0.99) RG (0.86)
[21]	X	X		X				X	X			X		RF (0.8938) RG (0.8854)

Table 4
Machine learning techniques used to train/test the models and description to reproduce the process. “Class imbalance considered?”: whether the paper described how class imbalance was handled when training the models. “Dataset split into train/test sets”: whether the paper described how the dataset was split into a training and a test set. “Hyperparameters described?”: whether the paper described all the hyperparameters used to train the models (“partial” means some, but not all hyperparameters were described). “Version of tools and libraries listed?”: whether the paper listed the version of all tools and libraries used to train the models (“partial” means the version for some of the tools or libraries were listed, but not all). “Dataset available?”: whether the paper made the dataset available to train the models (after preprocessing) or the tools to create the dataset from the original source. “Code available?”: whether the paper made available the code to repeat the model training.

Paper	Class imbalance considered?	Dataset split into train/test sets?	Hyperparameters described?	Version of tools and libraries listed?	Dataset available?	Code available?
[9]	Not described	Split by unspecified method	No	No	No	No
[10]	Not described	Did not split	No	No	No	No
[11]	Not described	Split by unspecified method	Yes	Partial	Yes	Yes
[12]	Not balanced	Split by random assignment	No	No	No (1)	Yes
[13]	Oversampling	Not described	Partial	Partial	No	No
[14]	Not described	Split by year	No	No	No	No
[15]	Unspecified method	Split with stratification	Yes	No	No	No
[6]	SMOTE	Split by class prevalence	No	No	No	No
[8]	Not described	Split by population and class prevalence	No	No	No (1)	No
[16]	Not described	Split by class prevalence	Yes	Partial	No	No
[17]	Not described	Split by unspecified method	Partial	Partial	No	Yes
[18]	Not described	Split by random assignment	Yes	Partial	Yes (2)	No
[19]	Oversampling	Split by unspecified method	Partial	Partial	No	No
[20]	Unspecified method	Split by class prevalence	Yes	Partial	No	No
[7]	Downsampling and SMOTE	Split by unspecified method	Partial	No	Yes (3)	Yes (3)
[21]	Downsampling	Did not split	No	No	No	No

SMOTE: Synthetic Minority Oversampling Technique is a technique to create samples for the minority class close to the feature space.
(1) Institution restricts access to the data. (2) Upon request, two years after publication. (3) Inquire the authors.

Table 5
Essential items for reproducible machine learning studies. Studies that do not report these items should be considered potentially non-reproducible and, therefore, less beneficial for the research community.

Item
Is the dataset available (subject to ethical and privacy considerations)?
Is the source code for dataset preprocessing, model training, and model evaluation available?
Are seeds for random number generators set?
Is the handling of missing data clearly explained?
Are the criteria to split the dataset into training, test, and validation sets clearly defined?
Are the hyperparameters selection and optimization defined?
Is the version of each tool, library, and framework described?

4.1.2. Interpretable vs. non-interpretable models

Using ROC AUC as a metric (an imperfect metric for this case [30], but the most-used metric across the papers), the rightmost column in table 3 shows that when an interpretable paper is the second-best model, it is not far off from the best model [6–8,12,14,19,21]. Given the high stakes in healthcare applications, we should favor interpretable models when their performance is sufficient for the task [31].

Future studies could investigate this trend and confirm that interpretable models such as logistic regression or decision trees perform well enough to avoid deploying non-interpretable models in this critical healthcare area. If a non-interpretable model must be deployed for OUD prediction, the study should define what reasons justify its use.

4.1.3. Reproducibility of experiments and results

As reported in other papers [32–34], the reproducibility of papers using ML for healthcare applications continues to be lacking. While publishing datasets in healthcare research can be hindered by privacy laws and ethical considerations, there is no excuse not to follow ML practices to allow the reproduction of the dataset preprocessing step and the model training and evaluation process. Ideally, the code should be publicly available.

The lack of good ML reproducibility practices in the papers makes it impossible to verify their claims (as [35] states, "[*r*]eproducibility failures don't mean a claim is wrong, just that evidence presented falls short of the accepted standard or that the claim only holds in a narrower set of circumstances than asserted.") The papers would be more valuable to the research community and their purported application if they followed good practices of ML reproducibility so that their claims could be verified, allowing them to be used as a solid base for future work.

Transparent and reproducible practices have been getting attention in the machine learning community [26–28,36–38]. Table 5 lists the essential items for transparent and reproducible machine learning studies, compiled from the references. We recommend that, at a minimum, reviewers request these items to be documented before accepting ML-based studies for publication. Better documented and reproducible studies will help the research community advance the field.

4.2. Limitations

This review has the following limitations:

(1) It primarily analyzes papers published in healthcare journals. They may be skewed toward inexperienced ML authors and peer reviewers compared to papers published in computer science journals.
(2) It did not verify if the models were tried in clinical applications or other settings for which they were created. We could not find evidence that they were tried in practice, but we have not formally reviewed this aspect.
(3) It covers a subset of opioid use disorder research, namely the prediction of OUD. ML models are being applied to other areas of the opioid crisis, such as overdose prediction. Those areas could benefit from a similar review investigating their results and use of ML practices.
(4) It uses ROC AUC to compare models, an imperfect metric for imbalanced datasets where the negative class is significantly larger than the positive class [30]. To mitigate this limitation, we avoided directly comparing the papers and compared interpretable and non-interpretable models within the same paper. Future papers should use better metrics for imbalanced datasets in healthcare applications, such as AUPRC.
(5) It did not analyze the predictors of the machine learning models selected within the same paper (by different models) and across papers. This analysis can provide more insights into what the papers identify as predictors for OUD.
(6) It did not check for data leakage that may happen with datasets that cover multiple years (training with the full range of years instead of reserving the most recent years for the test set).

Future systematic reviews can help advance the area by analyzing these items.

5. Conclusions

We reviewed 16 papers that use machine learning (ML) models to predict OUD. In addition to the final results of the models, we reviewed how the papers trained and evaluated the models. To our knowledge, this is the first systematic review that analyzes the technical aspects of machine learning applied to OUD prediction.

While the results from the reviewed papers indicate that ML models applied to OUD prediction may be useful, the lack of details and transparency in preprocessing the dataset, training, and evaluating the models hinders their application in real-life conditions and limits their use for research. In the "Challenges and recommendations" section, we list recommendations to improve future research on this topic.

Sources of funding

None.

Declaration of Competing Interest

The author of the manuscript "Machine learning for predicting opioid use disorder from healthcare data: a systematic review" declare that they have no conflict of interest to report for this work.

Appendix A

This appendix describes the queries used in each database.

Semantic Scholar

Semantic Scholar does not have a query language. We performed the following text searches to find relevant records. Key terms were quoted to reduce the number of unrelated items during searches.

(1) artificial intelligence opioid addiction prediction
(2) artificial intelligence "pain medication" addiction prediction
(3) artificial intelligence opioid abuse prediction
(4) artificial intelligence opioid dependence prediction
(5) artificial intelligence opioid misuse prediction
(6) artificial intelligence opioid "use disorder" prediction
(7) machine learning opioid addiction prediction
(8) machine learning "pain medication" addiction prediction
(9) machine learning opioid abuse prediction
(10) machine learning opioid dependence prediction
(11) machine learning opioid misuse prediction
(12) machine learning opioid "use disorder" prediction

We applied the following filters on the search web page:

- Date range: 2012 to 2022
- Fields of study: computer science, medicine, and sociology. The first two fields are obvious. We added the third one, sociology, because opioid use disorder is also a socioeconomic problem and therefore that field could have relevant material for our work.

Google Scholar

We used the following query and applied the date filter (2012 to 2022) on the search web page. allintitle:
(abuse OR dependence OR abuse OR misuse OR disorder)
("machine learning" OR "artificial intelligence")
(opioid OR opiate)

Science.gov

We used the query below, applied the data filter (2012 to 2022), and selected the categories "applied science & technologies: biotechnology, electronics, engineering, transport", "general science - multidisciplinary resources", "health & medicine - disease, health care, nutrition, mental health", and "public access - peer-reviewed scholarly publications resulting from federally funded scientific research" on the search web page.
("artificial intelligence" OR "machine learning")
AND (
(opioid OR opiate)
AND (addiction OR dependence OR abuse OR misuse OR disorder)
AND predict*
)

IEEE Xplore

We used the following query and applied the date filter (2012 to 2022) on the search web page.
(("Full Text & Metadata":"artificial intelligence") OR ("Full Text & Metadata":"machine learning"))
AND (
(("Full Text & Metadata":opioid) OR ("Full Text & Metadata":opiate))
AND (
("Full Text & Metadata":addiction) OR
("Full Text & Metadata":abuse) OR
("Full Text & Metadata":misuse) OR
("Full Text & Metadata":disorder)
)
AND
("Full Text & Metadata":predict*)
)

PubMed

We used the following query, which includes the years as a filter:
"(artificial intelligence[tw] OR machine learning[*tw*])
AND (
(opioid[tw] OR opiate[tw)
AND (addiction OR dependence OR abuse OR misuse OR disorder)
AND predict*
)
AND (("2012″[Date - Publication] : "2022″[D*ate* - P*ublication*]))

Appendix B

This appendix describes the process to select the papers.

First, duplicated items were merged using Zotero's duplicate detection functionality. The authors found a small number of duplicate items not identified by Zotero and manually removed them during the review phase.

After duplication removal, two researchers (NM, CG) worked together in the following phases to screen the search results. Each phase acted as a filter, removing items before proceeding to the next phase. In each phase, the decision of each researcher was recorded as a tag in Zotero. With the tags in place, the researchers used Zotero's tag filtering to identify each other's decisions and discuss the differences.

(1) Title review: The researchers independently chose "keep" or "remove" for each item, based on the item's title, then met to resolve differences by consensus.
(2) Keywords and abstract review: The researchers independently chose "keep" or "remove" for each item, based on the item's keywords (if present) and abstract, then met to resolve differences by consensus.
(3) Article type review: The researchers worked independently to remove the non-peer-reviewed items, then met to review the list together.
(4) Full-text review: The researchers independently retrieved the full text for each item and chose "keep" or "remove" for them. Each researcher reviewed half of the items still left up to this point. Articles were split by first author's last name to reduce the chances that one reviewer ended up with all articles of specific types (which could happen if the list was split by item title). They met to explain their individual reasons to keep or remove, resolving differences by consensus.

The reasons and number of papers to include papers during the full-text review were as follows:

- Review papers ($n = 2$).
- About other drugs, not opioids ($n = 3$).
- Used social media data mining, not healthcare data ($n = 2$).
- Legitimate use of opioids, not abuse ($n = 6$).
- Comparison of ML models, not creating new models ($n = 1$).
- Survival analysis, not prediction ($n = 2$).

Appendix C

The following tables expand on Table 3, showing in more detail the machine learning models (C.1) and metrics (C.2) used in each paper.

Table C.1
What machine learning models the papers used.

Paper	Regression					DT	RF	GB	KNN	SVM	MLP, DNN	CNN	RNN	TF
	LR1	LR2	EN	LRNA	COX									
[9]			X											
[10]										X				
[11]								X						
[12]				X		X	X				X		X	X
[13]							X							
[14]				X		X	X			X	X			
[15]					X		X	X			X			
[6]				X		X	X	X						
[8]			X				X	X			X			
[16]		X												X
[17]	X						X		X	X				
[18]								X						
[19]	X										X	X		
[20]			X											
[7]		X						X						
[21]				X		X	X							

LR1: logistic regression (LR) LASSO, LR2: LR Ridge, EN: LR ElasticNet, LRNA: LR without regularization or not described, COX: Cox regression, DT: decision tree, RF: random forest, GB: gradient boosting, including XGBoost, KNN: k-nearest neighbor, SVM: support vector machine, MLP: multi-layer perceptron, a.k.a fully connected networks, DNN-deep neural network; CNN: convolutional neural network, RNN: recurrent neural network (including LSTM), TF: transformer-based network.

Table C.2
What metrics the papers used to evaluate the ML models.

Paper	Specificity, TNR	Sensitivity, recall, TPR	Precision, PPV	NPV	ROC AUC value, c-statistic (1)	ROC AUC graph	AUPRC value	AUPRC graph
[9]					X	X		
[10]					X	X		
[11]			X		X			
[12]		X	X		X	X		
[13]	X	X	X	X	X	X		
[14]			X		X			
[15]					X	X	X	X
[6]		X	X		X	X		
[8]	X	X	X	X	X	X	X	X
[16]		X	X					
[17]	X	X	X	X	X			
[18]	X	X	X	X	X	X	X	X
[19]	X	X	X	X	X			
[20]					X		X	
[7]		X	X	X	X	X	X	X
[21]	X	X			X			

PPV: positive predictive value, NPV: negative predictive value, TPR: true positive rate, TNR: true negative rate, ROC AUC: receiver operating characteristic area under the curve, AUPRC: area under the precision-recall curve. Note that NNE (number needed to evaluate) can be calculated from PPV, thus it is not reported separately. Similarly, the F1 score can be calculated from precision and recall, thus not reported separately.
(1) For binary outcomes, the c-statistic and the ROC AUC are the same, thus reported in the same column in this table.

Appendix D

Table D.1 shows the number of papers found with the database searches, before filtering the results. The intention of this table is to verify if this review missed older papers since the filtering process identified 13 of the 16 papers as published in 2020 or later. The table indicates that there has been an uptick in papers in recent years in this area. That is evidence that our filtering process did not unduly remove older papers.

Table D.1
The number of papers found with the database searches before any filtering, by year of publication.

Year	Number of papers
2012	10
2013	21
2014	21
2015	24
2016	28
2017	60
2018	101
2019	170
2020	266
2021	285
2022	239

References

[1] Centers for Disease Control and Prevention, “Opioid Data Analysis and Resources | Opioids | CDC,” Jun. 01, 2022. https://www.cdc.gov/opioids/data/analysis-resources.html (accessed Nov. 04, 2022).
[2] U.S. Department of Health and Human Services, “What is the U.S. Opioid Epidemic?,” HHS.gov, Dec. 04, 2017. https://www.hhs.gov/opioids/about-the-epidemic/index.html (accessed Nov. 03, 2022).
[3] K.K. Mak, K. Lee, C. Park, Applications of machine learning in addiction studies: a systematic review, Psychiatry Res. 275 (2019) 53–60 May, doi:10.1016/j.psychres.2019.03.001.
[4] S. Gadhia, G.C. Richards, T. Marriott, J. Rose, Artificial intelligence and opioid use: a narrative review, BMJ Innov. 9 (2) (2023) Apr., doi:10.1136/bmjinnov-2022-000972.
[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, Curran Associates, Inc., 2012 Accessed: Nov. 01, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[6] M.M. Hasan, G.J. Young, M.R. Patel, A.S. Modestino, L.D. Sanchez, M.D. Noor-E-Alam, A machine learning framework to predict the risk of opioid use disorder, Mach. Learn. Appl. 6 (2021) 100144 Dec, doi:10.1016/j.mlwa.2021.100144.
[7] R. Vunikili, B. Glicksberg, K.W. Johnson, J. Dudley, L. Subramanian, K. Shameer, Predictive modeling of susceptibility to substance abuse, mortality and drug-drug interactions in opioid patients, Front. Artif. Intell. (2021) Dec., doi:10.3389/frai.2021.742723.
[8] W. Lo-Ciganic, et al., Using machine learning to predict risk of incident opioid use disorder among fee-for-service Medicare beneficiaries: a prognostic study, PLoS One (2020) Jul., doi:10.1371/journal.pone.0235981.
[9] W.Y. Ahn, J. Vassileva, Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence, Drug Alcohol Depend. 161 (2016) 247–257 Apr., doi:10.1016/j.drugalcdep.2016.02.008.
[10] L.A. Averill, C.L. Averill, L.A. Staley, J. Ozawa-Kirk, J.S. Kauwe, P. Henrie-Barrus, The opioid abuse risk screener predicts aberrant same-day urine drug tests and 1-year controlled substance database checks: a brief report, Health Psychol. Open 4 (2) (2017), doi:10.1177/2055102917748459. Art. no. 2 Jul.
[11] O. Corradin, et al., Convergence of case-specific epigenetic alterations identify a confluence of genetic vulnerabilities tied to opioid overdose, Mol. Psychiatry 27 (4) (2022), doi:10.1038/s41380-022-01477-y. Art. no. 4 Apr.
[12] X. Dong, et al., Identifying risk of opioid use disorder for patients taking opioid medications with deep learning, J Am Med. Inform. Assoc. 28 (8) (2021) 1683–1693 Apr., doi:10.1093/jamia/ocab043.
[13] R.J. Ellis, Z. Wang, N. Genes, A. Ma'ayan, Predicting opioid dependence from electronic health records with machine learning, BioData Min. (2019) Jan., doi:10.1186/s13040-019-0193-0.
[14] W. Gao, C. Leighton, Y. Chen, J. Jones, P. Mistry, Predicting opioid use disorder and associated risk factors in a medicaid managed care population, Am. J. Manag. Care 27 (4) (2021), doi:10.37765/ajmc.2021.88617. Art. no. 4 Apr.
[15] D.H. Han, S. Lee, D. Seo, Using machine learning to predict opioid misuse among U.S. adolescents, Prev. Med. 130 (2020) Jan, doi:10.1016/j.ypmed.2019.105886.
[16] M. Poulsen, P. Freda, V. Troiani, A. Davoudi, D. Mowery, Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing, Front. Public Health (2022) May, doi:10.3389/fpubh.2022.850619.
[17] J.T. Prieto, et al., The detection of opioid misuse and heroin use from paramedic response documentation: machine learning for improved surveillance, J. Med. Internet Res. 22 (1) (2020) Jan., doi:10.2196/15645.
[18] Z. Segal, et al., Development of a machine learning algorithm for early detection of opioid use disorder, Pharmacol. Res. Perspect. 8 (6) (2020) Nov., doi:10.1002/prp2.669.
[19] B. Sharma, et al., Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients, BMC Med. Inform. Decis. Mak. (2020) Apr., doi:10.1186/s12911-020-1099-y.
[20] C.L. Thompson, K.C. Alcover, S.W. Yip, Clinical prediction of extra-medical use of prescription pain relievers from a representative United States sample, Prev. Med. 149 (2021) 106610 Aug., doi:10.1016/j.ypmed.2021.106610.
[21] A.S. Wadekar, Understanding opioid use disorder (OUD) using tree-based classifiers, Drug Alcohol Depend. 208 (2020) 107839 Mar, doi:10.1016/j.drugalcdep.2020.107839.
[22] K. Dickersin, Y.I. Min, Publication bias: the problem that won't go away, Ann. N. Y. Acad. Sci. 703 (1993) 135–146 discussion 146-148Dec., doi:10.1111/j.1749-6632.1993.tb26343.x.
[23] J.D. Scargle, "Publication Bias (The 'File-Drawer Problem') in Scientific Inference." arXiv, Sep. 17, 1999. doi:10.48550/arXiv.physics/9909033.
[24] SAMHSA (Substance Abuse and Mental Health Services Administration), "2020 NSDUH Annual National Report |CBHSQ Data," 2020. https://www.samhsa.gov/data/report/2020-nsduh-annual-national-report (accessed Nov. 02,2022).
[25] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (1) (2019) 27 Mar., doi:10.1186/s40537-019-0192-5.
[26] A. Ng and K. Katanforoosh, "Splitting into train, dev and test sets," 2022. https://cs230.stanford.edu/blog/split/ (accessed Nov. 02, 2022).
[27] O.E. Gundersen, S. Kjensmo, State of the Art: reproducibility in Artificial Intelligence, in: Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018, doi:10.1609/aaai.v32i1.11503. Art. no. 1 Apr.
[28] S. Kakarmath, et al., Best practices for authors of healthcare-related artificial intelligence manuscripts, Npj Digit. Med. 3 (1) (2020), doi:10.1038/s41746-020-00336-w. Art. no. 1 Oct.
[29] N. Barnes, Publish your computer code: it is good enough, Nature 467 (7317) (2010) Art. no. 7317 Oct., doi:10.1038/467753a. Art. no. 7317 Oct.
[30] T. Saito, M. Rehmsmeier, The Precision-recall plot is more informative than the ROC Plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (3) (2015) e0118432 Mar, doi:10.1371/journal.pone.0118432.
[31] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) Art. no. 5May, doi:10.1038/s42256-019-0048-x. Art. no. 5 May.
[32] C.L. Andaur Navarro, et al., Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review, BMC Med. Res. Methodol. 22 (1) (2022) 12 Dec., doi:10.1186/s12874-021-01469-6.
[33] M.B.A. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, M. Ghassemi, Reproducibility in machine learning for health research: still a ways to go, Sci. Transl. Med. 13 (586) (2021) eabb1655 Mar., doi:10.1126/scitranslmed.abb1655.
[34] A.L. Beam, A.K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, JAMA 323 (4) (2020) 305–306 Jan., doi:10.1001/jama.2019.20866.
[35] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in ML-based Science." arXiv, Jul. 14, 2022.Accessed: Nov. 02, 2022. [Online]. Available: http://arxiv.org/abs/2207.07048.
[36] NeurIPS, "NeurIPS 2021 Code and Data Submission Guidelines," 2021. https://nips.cc/Conferences/2021/PaperInformation/CodeSubmissionPolicy (accessed Apr. 14, 2023).
[37] J. Pineau et al., "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019Reproducibility Program)." arXiv, Dec. 30, 2020. Accessed: Apr. 13, 2023. [Online]. Available: http://arxiv.org/abs/2003.12206.
[38] Association for the Advancement of Artificial Intelligence, "Reproducibility Checklist," AAAI. https://aaai.org/conference/aaai/aaai-23/reproducibility-checklist/ (accessed Apr. 14, 2023).