Development and External Validation of ML Models for Identifying Patients at Risk of Postoperative Prolonged Opioid Use

A Network Study on OMOP Databases

Stanford MEDICINE | Center for Biomedical Informatics Research

Study Protocol

Document Contributors

Institution/Affiliation	Author	Email
Stanford University	Behzad Naderalvojoud	behzadn@stanford.edu
Stanford University	Catherine Curtin	boussard@stanford.edu
Stanford University	Tina Hernandez-Boussard	ccurtin@stanford.edu

Document History

Date	Version	Contributors	Status
08/01/2022	0.1	BN	Table of Content
08/28/2022	0.8	BN	First draft
09/05/2022	0.9	BN, CC, THB	Review
09/10/2022	1.0	BN	Final draft
12/20/2022	1.1	BN	Adding project collaborators
03/01/2023	2.0	BN and all collaborators	Adding new experiment settings

Table of Contents

List of Figures

List of Tables

Glossary

Acronym	Meaning
PORPOISE	POstopeRative Prolonged OpioId uSE
OMOP	Observational Medical Outcomes Partnership
CDM	Common Data model
NLM	National Library of Medicine
IRB	Institutional Review Board
ML	Machine Learning
PLP	Patient Level Prediction
PPV	Positive Predictive Value
NPV	Negative Predictive Value
ROC	Receiver Operating Characteristics
PRC	Precision Recall Curve
AUC	Area Under Curve
LLR	Lasso Logistic Regression
RF	Random Forest
AB	AdaBoost
GBM	Gradian Boosting Machine
NB	Naive Bayes

1 Project basics

1.1 Sponsor

The project is supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013362 and has received approval from the Institutional Review Board (IRB) at Stanford University.

1.2 Responsible parties

Principle Investigator:

Tina Hernandez-Boussard, PhD MPH, MS, FACMI
Stanford University Biomedical Informatics Research (BMIR), USA

Study lead:

Behzad Naderalvojoud, PhD
Stanford University Biomedical Informatics Research (BMIR), USA

Table 1. PORPOISE collaborators.

Name	Affiliation
Catherine Curtin	Stanford Medicine, USA
Tina Seto	Stanford University, USA
Chen Yanover	KI Research Institute, Israel
Tal El-Hay	KI Research Institute, Israel
Byungjin Choi	AJOU University School of Medicine, South Korea
Rae Woong Park	AJOU University School of Medicine, South Korea
Javier Gracia Tabuenca	Helsinki University, Finland
Mary Pat Reeve	Helsinki University, Finland
Thomas Falconer	Columbia University, USA

1.3 Description

Opioids are potent analgesics often used to manage pain, including postoperative pain. However, opioids can be highly addictive, even when prescribed correctly and taken as directed. The balance between pain management and opioid misuse is challenging. To improve patient outcomes following surgery, it is crucial to identify patients at risk for prolonged opioid use prior to prescribing pain management regimes. Many studies have identified a limited number of features predictive of prolonged postoperative opioid use from real-world data, but these are isolated studies using non-standardized data from different sources, which limits their generalizability and reliability across populations. In this study, we develop and validate five machine learning (ML) algorithms to predict patients at risk of prolonged opioid use in a diverse, multisite cohort study by evaluating not only the performance, but also their discrimination and calibration abilities. We address generalizability limitations by using the Observational Medical

Outcomes Partnership (OMOP) Common Data Model (CDM) to identify postoperative patients at higher risk of prolonged opioid use based on preoperative risk factors using ML approaches.

1.4 Goals

The project main goals are classified into clinical and methodological goals as follows:

- Clinical goals:
  - Improve pain management following surgery.
  - Identify patients at risk for prolonged opioid use prior to prescribing pain management regimens.
  - Identify common preoperative risk factors predictive of postoperative opioid use over CDM databases.
- Methodological goals:
  - Develop and validate ML models in a diverse, multisite cohort study by evaluating their generalizability, discrimination, and calibration abilities.
  - Evaluate the generalizability and calibration of the ML models across multisite cohort subgroups: diabetes, depression, obesity.
  - Evaluate the transportability of ML models based on population differences in the various CDM databases.
  - Incorporate ML models trained on different databases to increase generalizability.
  - Incorporating developed ML models into an open-access web application so that researchers can evaluate the performance of models and compare them to their own models in a transparent setting.

2 Rationale and background

Opioids are potent analgesics often used to manage pain, including postoperative pain. However, opioids can be highly addictive, even when prescribed correctly and taken as directed. Over the past 15 years, the number of opioid-related drug overdose deaths has tripled, with prescription opioids accounting for more than half of all overdose deaths (Dowell et al, 2016). On the other hand, opioid use has been linked to serious complications such as dependence and abuse, which have resulted in significant morbidity and mortality (Lyden & Binswanger, 2019). Although postoperative opioid exposure is a major risk factor for prolonged use and abuse, prescription opioid medications continue to play important roles in the management of postoperative pain following surgery (Brummett et al, 2017). To reduce the morbidity associated with opioid use after surgery, it is crucial to identify patients at risk for prolonged opioid use prior to prescribing opioids. Several studies have used machine learning (ML) to predict patients at risk and to identify risk factors associated with this complication for specific surgeries, such as arthroscopic meniscectomy and spine surgery (Karhade et al, 2020, Lu et al, 2022). In this study, we will develop and validate ML models to predict patients at risk of prolonged opioid use for most surgeries. Brummett et al. found no difference in new persistent opioid use between patients who underwent minor and major surgical procedures, indicating that prolonged opioid use is not solely caused by surgical pain. Therefore, in developing ML models, we will consider patients' 6-month prior medications, diagnoses, lab measurements, and procedures.

Moreover, many patient-specific clinical parameters complicate the prediction of prolonged opioid use. Many studies have identified a limited number of features predictive of prolonged

postoperative opioid use from real-world data, but these are isolated studies using non-standardized data from different sources, which limits their generalizability and reliability across population (Katakam et al, 2020; Dong et al, 2021; Ward et al, 2021). This study aims to address these limitations, by using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to identify postoperative patients at higher risk of prolonged opioid use based on preoperative risk factors using ML approaches. The OMOP CDM, developed by the Observational Health Data Sciences and Informatics (OHDSI) community, standardizes the format of observational healthcare data, allowing the analysis code to be directly shared with other researchers working with different datasets. We will also investigate the impact of CDM features on the performance of several ML models using various feature evaluation metrics and provided transparent validation on model discrimination, calibration, and clinical utility. Unlike previous studies, which only used a limited number of features determined in the literature, using CDM features allows us to consider a wide range of covariates across domains, such as demographic, drug, condition, procedure, and measurement, in the development of ML models without the need for manual feature engineering.

3 Materials and methods

3.1 Data source

The study will rely on multisite observational data from electronic health records (EHRs) mapped to the OMOP CDM. All the data will be analyzed in a federated manner, where the data will remain with the data owners and only the analysis results will be shared and published. The datasets used in the PORPOISE study are summarized in Table.

Table 2. PORPOISE data sets.

CDM Dataset	Dataset Name	Country
STARR-US	STAnford Research Repository OMOP	United States
IMRD-UK	IQVIA Medical Research Data UK	United Kingdom
FinnGen-FI	FinnGen data freeze 10	Finland
AUSOM-KR	Ajou University School of Medicine	South Korea
CUIMC-US	Columbia University Irving Medical Center	United States

STARR-OMOP: The STARR-OMOP is Stanford electronic health record data from its two hospitals in the OMOP Common Data Model. This dataset was created from Epic Clarity data and was normalized and deidentified by Research IT at Stanford Medicine for research purposes. It is now available to the Stanford research community.

IMRD-UK: IQVIA Medical Research Data (IMRD) incorporates data from The Health Improvement Network (THIN), a Cegedim database[1]. Cegedim provides access to precise, ethical, actionable, and anonymized longitudinal patient data from the THIN database. It includes records collected from patient management software used by general practitioners. Reference made to THIN is intended to be descriptive of the data asset licensed by IQVIA.

[1] https://www.cegedim-health-data.com/cegedim-health-data/thin-the-health-improvement-network/

FinnGen: The FinnGen[2] dataset was collected through FinnGen project launched in 2017. This project is a public-private research project, combining genome and digital healthcare data on eventually 500,000 Finns. The nation-wide research project aims to provide novel medically and therapeutically relevant insight into human diseases. FinnGen is a pre-competitive partnership of Finnish biobanks and their background organizations (universities and university hospitals) and international pharmaceutical industry partners and Finnish biobank cooperative (FINBB). Register data includes diagnoses and operations from primary and specialist healthcare visits, pharmacy drug purchases, cancer register, and death registry.

AUSOM: The AUSOM dataset contains electronic health record data from Ajou University School of Medicine, a tertiary teaching hospital in Korea.

3.2 Study design

The PORPOISE study is a retrospective analysis of observational health data that has received approval from the Institutional Review Board (IRB) at Stanford University. The analysis and prediction models were implemented using the OHDSI PatientLevelPrediction (PLP) package,[3] and all study materials are available in the project GitHub repository to be run on any OMOP databases[4]. In the study, all observations about opioid medications are looked at in two different time periods: the preoperative period, which is a six-month observation window before surgery, and a six-month follow-up period with a time-at-risk (TAR) between 90 and 180 days after surgery. Based on these two time periods, the following analyses are performed:

- The target and outcome cohorts, as well as evaluation subgroups, are defined.
- The cohorts included in the study are characterized based on the predefined features across three subgroups: diabetes, depression, and obesity.
- Different ML algorithms are trained using different sets of covariates and validated over the target cohort and three risk subgroups.
- The trained models are shared across partners for external validation.
- Prolonged opioid use risk factors that are common across countries are identified.
- All internal and external validation results are discussed and published.

The experiments are conducted to answer the following questions:

1) What is the impact of ML algorithms and covariates for developing generalizable models for predicting postoperative prolonged opioid users?
2) Which is more important in generalizability: covariates or ML algorithms?
3) Are ML models generalizable across high-risk subgroups such as diabetes, depression, and obesity?
4) What is the performance of US-developed models versus European and Asian-developed models with different opioid policies?
5) What are the risk factors that are common across countries?

[2] https://www.finngen.fi/en

[3] https://github.com/OHDSI/PatientLevelPrediction

[4] https://github.com/ohdsi-studies/PORPOISE

To answer these questions, we develop ten models using five ML algorithms, including Lasso Logistic Regression (LR), Random Forest (RF), AdaBoost (AB), Extreme Gradient Boost (GB), and Naive Bayes (NB), and two sets of covariates, as follows:

- The most relevant covariates predictive of prolonged opioid use, as determined by the chi-square metric and significant at P-value < 0.01 from the development dataset,
- Superset (union) of covariates selected by the chi-square metric from all training sets used in development and external validation databases.

The models trained by features selected from development database are validated both internally and externally across the target cohort and three high-risk groups as follows:

- **Internal validation:** 20% of the target cohort selected as a test set in the development database, stratified by three high-risk groups.
- **External validation:** 20% of the target, diabetes, depression, and obesity cohorts selected as a test set in the external validation databases.
- **External validation:** 100% of the target, diabetes, depression, and obesity cohorts in the external validation databases.

The models trained by a superset of features selected from development and validation databases are validated both internally and externally across the target cohort and three high-risk groups as follows:

- **Internal validation:** 20% of the target cohort selected as a test set in the development database, stratified by three high-risk groups.
- **External validation:** 20% of the target, diabetes, depression, and obesity cohorts selected as a test set in the external validation databases.

To quantify the biases introduced by the follow-up inclusion criteria (the last three rules shown in Figure 2), we validate the external performance of the best-performing models twice on cohorts that include or do not include follow-up inclusion criteria.

To determine the risk factors, the intersection of the features selected by the PNF and chi-square metrics are used (Naderalvojoud & Sezer, 2020).

As a result, the project includes the following evaluation components:

- five evaluation cohorts:
    1. target,
    2. target with no post-criteria,
    3. diabetes,
    4. depression
    5. obesity.
- five evaluation cohorts, not including patients used in training.
- Two types of models:
    a) five models trained with internal features.
    b) five models with superset features.
- Three sets of external validation results:
    a) five models trained with internal features validated on 5 cohort evaluations.
    b) five models trained with internal features validated on 5 test set of cohort evaluations.
    c) five models trained with superset features validated on 5 test set of cohort evaluations.
- Two sets of internal validation results:

a) five models trained with internal features validated on 5 test set of cohort evaluations.
b) five models trained with superset features validated on 5 test set of cohort evaluations.

To assess the impact of US-developed models versus European and Asian models, we develop our models based on three databases, STARR-US, IMRD-UK, and AUSOM-KR and compare their generalizability with each other. The overall structure of the model development and evaluation is summarized in Figure 1.

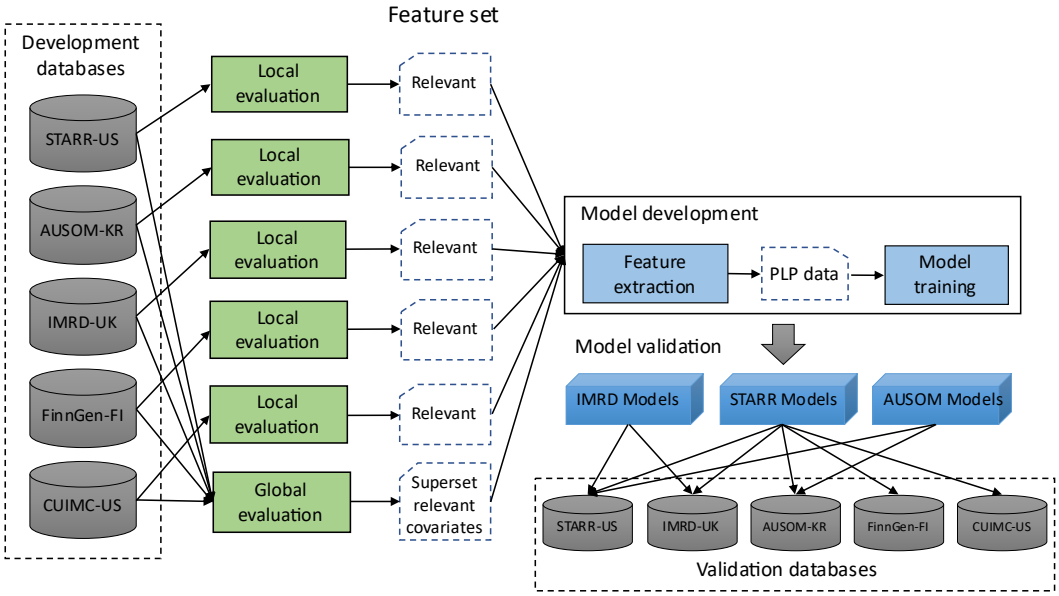

Figure 1. Overall structure of the model development and evaluation

3.2.1 Target cohort

The target cohort includes adult patients who underwent surgery during an inpatient visit between 2008 and 2019 with at least one opioid drug exposure 30 days before or after the surgery. Patients were included if they had at least two visits between two years and 30 days before the surgery and two visits within 30 days to two years after the surgery. If a patient had multiple surgeries, only the first event was included. We also excluded patients who had any other surgery from two months to seven months after the index surgery. Figure 2 outlines the inclusion criteria for the target cohort.

In the target cohort, nine groups of RxNorm opioid drug ingredients, shown in Table 3, are used to determine opioid prescriptions, and ICD/CPT codes associated with the following seventeen key groups of surgeries are considered to identify target procedures. As a result of ICD/CPT code mapping, 5,180 CDM concepts were employed to identify target procedures.

1) Laminectomy, excision intervertebral disc
2) Spinal fusion
3) Cholecystectomy and common duct exploration
4) Partial excision bone
5) Hysterectomy, abdominal and vaginal
6) Colorectal resection
7) Excision, lysis peritoneal adhesions

8) Appendectomy
9) Treatment, fracture or dislocation of hip and femur
10) Oophorectomy, unilateral and bilateral
11) Coronary artery bypass graft (CABG)
12) Inguinal hernia repair
13) Distal radial Fracture
14) Thoracotomy
15) Mastectomy
16) Knee Replacement
17) Treatment, fracture or dislocation of lower extremity (other than hip or femur)

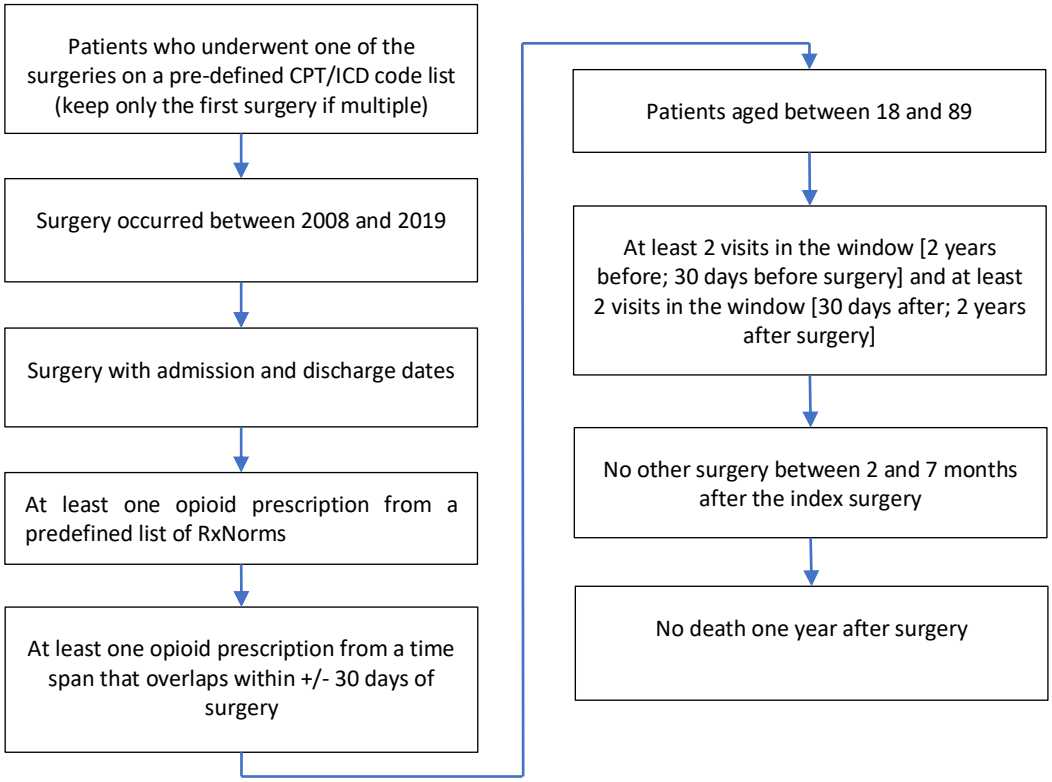

Figure 2. Target cohort inclusion criteria.

Table 3. RxNorms used to identify opioid prescriptions.

RxNorm	Drug ingredient
5489	Hydrocodone
4337	Fentanyl
2670	Codeine
3423	Hydromorphone
6754	Meperidine
6813	Methadone
7052	Morphine
7804	Oxycodone
10689	Tramadol

To characterize the study datasets and evaluate the performance of prediction models, three subgroups, including diabetes, depression, and obesity, are defined over the target cohort.

3.2.1.1 Diabetes subgroup

Occurring any of the following criteria is considered to determine diabetes subgroup:

1) At least one diagnosis of Type 1 or 2 diabetes mellitus all days before the index surgery, excluding diabetes mellitus during pregnancy.
2) At least one drug exposure, BLOOD GLUCOSE LOWERING DRUGS, EXCL. INSULINS, all days before the index surgery.
3) At least a measurement of Hemoglobin A1C with value between 6.5 and 20 percent all days before the index surgery.

3.2.1.2 Depression subgroup

To determine the depression subgroup, the following criterion is considered:

1) At least one occurrence of depressive disorder diagnoses 365 days before the index surgery, excluding postpartum depression.

3.2.1.3 Obesity subgroup

To determine the obesity subgroup, the following criterion is considered:

1) At least one occurrence of obesity diagnoses 180 days before the index surgery, excluding particular diagnoses such as Maternal Obesity Syndrome, Intellectual Disability, Seizures, Macrocephaly, Obesity Syndrome, Choroideremia With Deafness And Obesity Syndrome.
2) At least one occurrence of obesity observation 180 days before the index surgery, such as:
    a. Body mass index 30+ - obesity
    b. Body mass index 40+ - severely obese
    c. Obese class I (body mass index 30.0 - 34.9)
    d. Obese class II (body mass index 35.0 - 39.9)

3.2.2 Outcome

The primary outcome is prolonged opioid use, defined as an opioid drug exposure within three to six months after surgery. Based on this definition, two cohorts, namely prolonged and non-prolonged opioid users, are extracted from the target cohort to be used in characterization.

3.2.3 Cohort characterization

The four types of feature analysis, shown in Table 4, are used in the cohort characterization. The objective is to observe the distribution of features between prolonged and non-prolonged opioid users and understand the distribution differences between study datasets. Furthermore, three evaluation subgroups are used in the characterization to determine the distribution of these subgroups over individual features.

Table 4. Features used in characterization. Medium-term is considered the period beginning 180 days before the index date and ending on the index date.

Feature Type	Feature name
Demographics	Age
	Age group
	Ethnicity
	Race
	Gender
Distinct count	Distinct condition counts medium term
	Distinct procedure counts medium term
	Distinct measurement counts medium term
	Distinct ingredients count medium term
Clinical binary	Condition occurrence medium term
	Procedure occurrence medium term
	Measurement medium term
	Drug exposure medium term
Subgroup count	Diabetes count
	Depression count
	Obesity count

3.2.4 Patient Level Prediction

Figure 3 depicts the prediction problem based on a 180-day observation window and 90-day time-at-risk, as well as the study's outcome. As a result, the prediction problem is defined as estimating the risk of any opioid drug exposure at the time of risk based on patients' clinical attributes six months prior to surgery.

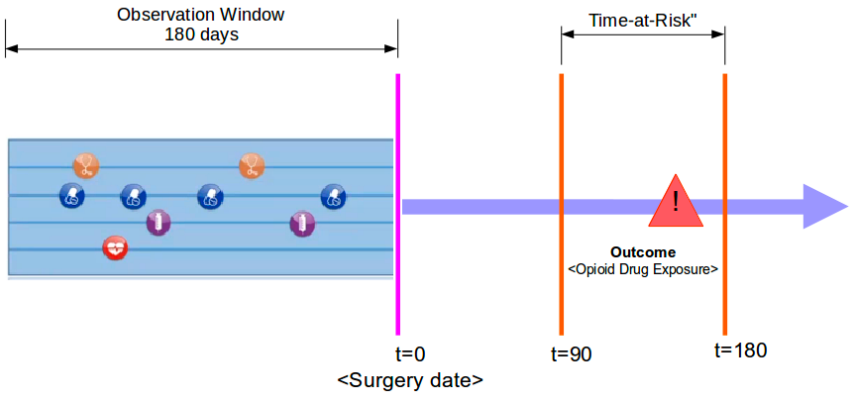

Figure 3. Prediction definition.

3.2.4.1 Covariate setting

All prediction features are defined based on CDM concepts shown in Table 5, and all infrequent covariates are removed with less than 0.001 frequency from the feature set.

Table 5. Prediction features.

Feature Type	Feature name
Demographics	Gender
	Age group
	Ethnicity
	Race
Distinct count	Distinct condition counts medium term
	Distinct procedure counts medium term
	Distinct measurement counts medium term
	Distinct ingredients count medium term
Clinical binary	Condition occurrence medium term
	Procedure occurrence medium term
	Measurement medium term
	Drug exposure medium term
Era and group binary	Drug era medium term
	Drug group era medium term
	Condition era medium term
Drug short term binary	Drug era start short term
	Drug exposure short term
	Distinct ingredients count short term

3.2.4.2 Machine leaning models

Five ML algorithms are used to develop our models: Lasso Logistic Regression (LR), Random Forest (RF), AdaBoost (AB), Gradian Boosting Machine (GBM), and Naive Bayes (NB). The models are developed using the PatientLevelPrediction (PLP) package from the OHDSI community, version 5.4.5. A stratified random sampling approach is used to split the dataset into train (80%) and test (20%) sets. To label patients in the target cohort, a 90-day time-at-risk (TAR) is defined between three and six months after surgery. For the ML models, any opioid drug exposure during this period is considered a positive case.

3.2.4.3 Internal evaluation

To evaluate models, standard metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are used. We also assess model discrimination using receiver operating characteristics (ROC) and precision recall curve (PRC) with their area under curve (AUC) values and 95% confidence intervals. Brier score and calibration curves are used to evaluate the model outputs in terms of their ability to generate calibrated probabilities for identifying patients at risk of prolonged opioid use. To select the best hyperparameters and prediction threshold, 5-fold cross validation on the train set is used with a grid search strategy.

3.2.4.4 External validation

After training and evaluating the models, the PLP results (the output of the prediction module) are shared with study partners to be validated externally on CDM databases where the models have not been trained. To that end, all models are validated on the entire target cohort as well as

the three subgroup cohorts on the target CDM database. All metrics used in the internal validation are adopted to evaluate external results. The distribution differences between the trained and validated databases will be used to discuss all external evaluations.

4 Dissemination plan

The study's findings will be presented as abstracts at Stanford AI + Health (2022) conference. The final results will be published in international peer-reviewed journals as full-text papers.

References

1. Datta, S., Posada, J., Olson, G., Li, W., O'Reilly, C., Balraj, D., ... & Shah, N. (2020). A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv preprint arXiv:2003.10534.
2. Dowell, D., Haegerich, T. M., & Chou, R. (2016). CDC guideline for prescribing opioids for chronic pain—United States, 2016. Jama, 315(15), 1624-1645.
3. Lyden, J., & Binswanger, I. A. (2019, April). The United States opioid epidemic. In Seminars in perinatology (Vol. 43, No. 3, pp. 123-131). WB Saunders.
4. Brummett, C. M., Waljee, J. F., Goesling, J., Moser, S., Lin, P., Englesbe, M. J., ... & Nallamothu, B. K. (2017). New persistent opioid use after minor and major surgical procedures in US adults. JAMA surgery, 152(6), e170504-e170504.
5. Karhade, A. V., Cha, T. D., Fogel, H. A., Hershman, S. H., Tobert, D. G., Schoenfeld, A. J., ... & Schwab, J. H. (2020). Predicting prolonged opioid prescriptions in opioid-naïve lumbar spine surgery patients. The Spine Journal, 20(6), 888-895.
6. Lu, Y., Forlenza, E., Wilbur, R. R., Lavoie-Gagne, O., Fu, M. C., Yanke, A. B., ... & Forsythe, B. (2022). Machine-learning model successfully predicts patients at risk for prolonged postoperative opioid use following elective knee arthroscopy. Knee Surgery, Sports Traumatology, Arthroscopy, 30(3), 762-772.
7. Dong, X., Deng, J., Rashidian, S., Abell-Hart, K., Hou, W., Rosenthal, R. N., ... & Wang, F. (2021). Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. Journal of the American Medical Informatics Association, 28(8), 1683-1693.
8. Katakam, A., Karhade, A. V., Schwab, J. H., Chen, A. F., & Bedair, H. S. (2020). Development and validation of machine learning algorithms for postoperative opioid prescriptions after TKA. Journal of orthopaedics, 22, 95-99.
9. Ward, A., Jani, T., De Souza, E., Scheinker, D., Bambos, N., & Anderson, T. A. (2021). Prediction of prolonged opioid use after surgery in adolescents: insights from machine learning. Anesthesia & Analgesia, 133(2), 304-313.
10. Weng, Y., Tian, L., Tedesco, D., Desai, K., Asch, S. M., Carroll, I., ... & Hernandez-Boussard, T. (2020). Trajectory analysis for postoperative pain using electronic health records: A nonparametric method with robust linear regression and K-medians cluster analysis. Health informatics journal, 26(2), 1404-1418.
11. Hah, J., & Hernandez-Boussard, T. (2018). Defining Postoperative Opioid Needs Among Preoperative Opioid Users. JAMA surgery, 153(7), 689-690.
12. Desai, K., Carroll, I., Asch, S. M., Seto, T., McDonald, K. M., Curtin, C., & Hernandez-Boussard, T. (2018). Utilization and effectiveness of multimodal discharge analgesia for postoperative pain management. Journal of Surgical Research, 228, 160-169.

13. Tedesco, D., Asch, S. M., Curtin, C., Hah, J., McDonald, K. M., Fantini, M. P., & Hernandez-Boussard, T. (2017). Opioid abuse and poisoning: trends in inpatient and emergency department discharges. Health Affairs, 36(10), 1748-1753.
14. Yim, W. W., Tedesco, D., Curtin, C., & Hernandez-Boussard, T. (2017, August). Annotation of pain and anesthesia events for surgery-related processes and outcomes extraction. In BioNLP 2017 (pp. 200-205).
15. Tedesco, D., Gori, D., Desai, K. R., Asch, S., Carroll, I. R., Curtin, C., ... & Hernandez-Boussard, T. (2017). Drug-free interventions to reduce pain or opioid consumption after total knee arthroplasty: a systematic review and meta-analysis. JAMA surgery, 152(10), e172872-e172872.
16. Chin, K. K., Desai, K., Asch, S. M., Carroll, I., Curtin, C., McDonald, K., & Hernandez-Boussard, T. (2017, June). Rapid Practice Changes in Response to the Opioid Epidemic: Integrating Adjuvant Analgesics into Practice. In 2017 Annual Research Meeting. AcademyHealth.
17. Naderalvojoud B, Sezer EA. (2020). Term evaluation metrics in imbalanced text categorization. Natural Language Engineering. 26(1), 31-47.