Mohl Jeff T (Orcid ID: 0000-0001-7415-9223)

Running title: Predicting Chronic Opioid Use

Title: Predicting chronic opioid use among patients with osteoarthritis using electronic health record data

Authors: Jeff T. Mohl, PhD*[1], Nikita Stempniewicz, MSc[1], John K. Cuddeback, MD, PhD[1], David M. Kent, MD, MS[2], Elizabeth A. MacLean, PharmD, PhD[3], Lance Nicholls, PharmD[3], Christopher Kerrigan, MD[4], Elizabeth L. Ciemins, PhD, MPH, MA[1]

Affiliations: [1]AMGA, Alexandria, VA. [2]Institute for Clinical Research and Health Policy Studies, Tufts-New England Medical Center, Boston, MA. [3]Pfizer Inc., New York, New York. [4]Community Memorial Hospital, Ventura, CA.

Financial Support: This study was sponsored by Pfizer

Disclosure: EAM and LN are employees of Pfizer with stock and/or stock options. AMGA, CK and DMK received funding from Pfizer for this study.

***Corresponding author:**

Email: jmohl@amga.org

Phone: 406-425-1097

Fax: 703-548-1890

Address: AMGA, 1 Prince street, Alexandria, VA, 22314

Word count: 3800

Other financial interests: Authors EAM and LN are employed by and own stock in Pfizer Inc.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/acr.25013

Abstract

Objective: To estimate the risk, using electronic health record (EHR) data and predictive models, of a patient with osteoarthritis (OA) developing chronic opioid use (COU) within 1 year of a new opioid prescription.

Methods: We used EHR data from 13 healthcare organizations to identify OA patients with an opioid prescription between 03/01/2017 and 02/28/2019 and no record of opioid use in the prior 6 months. We evaluated 4 machine learning models to estimate patients' risk of COU (≥3 prescriptions over ≥84 days, maximum gap ≤60 days). We also estimated the transportability of models to organizations outside the training set.

Results: The cohort consisted of 33,894 patients with OA, of whom 2,925 (8.6%) developed COU within 1 year. All models demonstrated good discrimination, with the best-performing model (random forest) achieving an area under the receiver operating characteristic curve (AUC) of 0.728 (95% CI, 0.711–0.745), but the simplest regression model performed nearly as well (AUC of 0.717; 95% CI, 0.699–0.734). Predicted risk deciles spanned a range of 2% risk for the 10th percentile to 18% risk for the 90th percentile for well-calibrated models. Models showed highly variable discrimination across organizations (AUC 0.571–0.842).

Conclusions: We found that EHR-based predictive models could estimate the risk of future COU among OA patients to help inform care decisions. Black-box methods did not have significant advantages over more interpretable models. Care should be taken when extending all models into organizations not included in model training due to high variability in performance across held-out organizations.

Significance and Innovation

- Used EHR data to predict whether a patient with OA will develop COU within 1 year of their first opioid prescription, allowing models to be integrated into the point of care to inform shared decision-making between patients and providers.
- Compared both interpretable (e.g., logistic regression) and black-box (e.g., random forest) models to determine whether improvements in performance justify loss of transparency.
- Leveraged a multi-organization dataset to evaluate model performance when applied to new organizations.

Osteoarthritis (OA) is the leading cause of chronic non-cancer pain, afflicting more than 32 million Americans in 2014 (1,2). While some patients may successfully control pain using over-the-counter analgesics or other therapies, estimates suggest that nearly a quarter of OA patients may be treated with prescription opioids in a given year (3). Chronic opioid use (COU) carries significant risks for patients, including the potential for opioid dependence, addiction, or overdose (4–6). In addition, patients often develop tolerance, requiring larger doses to maintain the same level of pain control. Consequently, initiation of chronic opioid therapy can be especially perilous because it may be challenging to later discontinue use. For these and other reasons, chronic use of opioids to treat chronic pain, including for patients with OA, has come under increasing scrutiny (7) and is recommended against in most circumstances (8). Therefore, tools for estimating the risk that a patient will develop COU prior to opioid initiation would assist in shared decision-making about treatment options for OA pain.

Clinical predictive models (CPMs), statistical techniques that leverage large datasets and computational approaches to identify trends and relationships regarding future health-related outcomes, may help to identify patients at elevated risk of COU. Similar models have been developed to predict a variety of outcomes, including opioid overdose (9,10), and clinical guidelines often incorporate such models (11–13), though to our knowledge none has been developed to predict COU among a population of primary care patients with OA.

We developed and evaluated a set of CPMs using electronic health record (EHR) data to estimate the risk that a given patient with OA, if prescribed an opioid, will experience a period of COU within 1 year. These models spanned a range of complexities, which allowed us to compare tradeoffs between model accuracy and interpretability and to recommend the best

modeling approach to balance these competing concerns. We also leveraged the multi-center nature of this dataset to evaluate the transportability of models to new health care organizations, providing an estimate of the range of expected model performance across settings.

Methods

Ethics statement

This study received IRB exemption status from WCG IRB.

Data source

EHR data (including outbound billing claims) from 13 geographically and EHR vendor-diverse AMGA (American Medical Group Association) member health care organizations (HCOs) were used. Data were extracted, mapped, and normalized by Optum® (Eden Prairie, MN); They contain up to 5.75 years of longitudinal clinical data (1/1/2015–9/31/2020) for individual patients receiving care within these multi-specialty medical groups and integrated health systems. Included patients reflected a broad and balanced representation of U.S. adults, e.g., by age, sex, race, ethnicity, health insurance type, and rural or urban residence. These data included patient demographics, encounters and procedures, diagnoses from outbound claims and patient problem lists, laboratory test results, clinical observations, and prescribed medications (e-prescriptions).

Study Design

We used a retrospective cohort design, following patients with OA who were prescribed an opioid without a recent (within prior 6 months) opioid prescription (Figure 1). The primary

outcome was whether the patient experienced a period of COU (defined below) in the 12 months following the initial prescription. Here, opioids refer only to orally administered opioid medications that are prescribed principally for pain treatment, not including those prescribed as part of an inpatient hospital visit, identified from electronic prescriptions in the EHR. This list included oral formulations of opioids and combinations with other medications (Table S1). Tramadol was excluded from this definition due to significant differences in pharmacology, provider attitudes, and prescribing patterns (7,8,14). Patients who experienced a period of COU were compared with patients in the same cohort who did not develop COU, in order to predict COU using EHR-derived variables.

The study population (Table 1) consisted of adult patients (ages 18–89) who received at least one opioid prescription during the index period (03/01/2017–02/28/2019), with no opioid prescription within the prior 6 months, and who had diagnosed OA on or before the index prescription date. OA was identified through the presence of two or more ambulatory visits with an OA diagnosis for any joint on a billing claim or if OA was on the patient's problem list, with diagnoses defined through ICD-9 and ICD-10 diagnosis codes (Table S2). Patients with no EHR activity (ambulatory encounters, clinical observations, lab results, or prescribed medications) at least 12 months prior to the index date were excluded to ensure adequate predictor ascertainment (Figure 1). We also required activity at least 3 months after the index date to exclude patients who did not have evidence of ongoing care following the initial prescription. Patients with a non-OA chronic pain condition for which opioids are commonly prescribed (diagnosis of malignant neoplasms, sickle cell disease, or cystic fibrosis during

baseline or follow-up), or those receiving palliative care or hospice services were also excluded. This resulted in a study population of 33,894 patients (Figure S1).

Definition of outcome

The primary clinical endpoint of interest was COU, here defined as at least one period of chronic use within 12 months of the index date. We defined a period of chronic use as any period with at least 3 opioid prescriptions, spanning at least 84 days with a maximum gap between consecutive prescriptions of ≤60 days. This definition was developed in a data-driven manner to closely match the most cited (15,16) definition of COU (90 days of supplied opioids, with <30-day gap in supply, using pharmacy claims data), while using only data that are reliably captured in the EHR. We validated this definition by examining a subset of patients with overlapping EHR and pharmacy claims data and then comparing our EHR-based definition with the claims-based definition (Table S3).

In our definition, prescriptions were excluded if they were prescribed within 3 days prior to or following an inpatient visit, observational stay, or surgical encounter of any kind. Prescriptions were also excluded if they could be clearly associated with an acute injury or other acute pain episode (defined by diagnosis codes on the same day as the index prescription, Table S2). This was done to maximize the likelihood that these prescriptions were prescribed for OA related pain, rather than unrelated conditions.

Predictive model development

We developed four separate models to predict risk of COU: logistic regression (LR), regularized logistic regression using elastic net (EN), support vector machine (SVM), and a random forest classifier (RF). Other than the differences in preprocessing described in the

following section, all models were tuned, trained, and evaluated on the same datasets, using the same validation procedure.

The dataset was randomly split into training (66%) and testing (34%) subsets, stratified on the outcome of COU. Within the training set, a folded cross-validation approach was used both for tuning and for internally validated estimates of performance throughout model development. Models were tuned using a 100-point, 5-fold pseudorandom grid search and the parameter values resulting in the highest mean area under the receiver operating characteristic curve (AUC) were selected for the final models. For internal cross-validation, a 10-fold, 3-repeat strategy was used to estimate model performance. The testing set was used for validation of the models after development was completed. This validation was separate from the non-random validation used in the transportability analysis (below).

Predictor selection and preprocessing

We extracted 92 features for each patient, including demographics, socioeconomic status, chronic condition diagnoses, prescriptions, clinical measures, non-pharmacological interventions, and healthcare utilization (Table S4). All features were captured from EHR data prior to or on the index date. Any feature with <1% prevalence in the population was eliminated to minimize overfitting to sparse features. Similarly, any categorical feature classes with <1% prevalence were aggregated into a single class. No other feature selection (e.g., univariate regression) was performed prior to model fitting, to minimize risk of bias (17).

In this dataset, only patient-reported pain scores had significant missingness (54.5%), while remaining features had a maximum missingness of <5%. Numerical missing values (other than pain score) were imputed using the median value from the remaining population, while

categorical missing values were assigned a factor level of "unknown" or similar as appropriate for each feature. This imputation impacted 8 variables and approximately 3% of patients. For pain scores, the high degree and non-random nature of missingness made imputation inappropriate. Because this variable was judged by the authors as likely relevant to the primary outcome, we chose to cast the numeric scale into 5 levels: mild (0–5), moderate (6–7), and severe (8–10) pain, missing at the individual level, or missing at the level of the HCO, i.e., no patients from the HCO in question had recorded pain scores. These levels were chosen to correspond with pain-related loss of function (18) and to capture a potentially important distinction between organizations that do not include pain scores in a structured field in the EHR and patients who could plausibly have had a pain score recorded but did not.

For modeling purposes, all categorical levels were encoded using dummy variables. For regression models, numeric variables were centered and scaled before model fitting. We found that class balancing using up or down-sampling, synthetic oversampling, or similar methods did not have an impact on model discrimination ability for most models, despite the relatively rare outcome. SVM was the exception, where class balancing was required to ensure the model could be fit properly. We therefore included a down-sampling step when fitting the SVM, while all other models were fit using the unmodified data subset.

Model Evaluation

Model performance was principally evaluated using AUC to compare the discrimination ability of each model. For purposes of calculating sensitivity and specificity, we tuned the decision threshold (what level of predicted risk corresponds with a "chronic" prediction) to maximize F-2 score for each model individually (19). F-score is a harmonic mean of sensitivity

and precision; it can be used to select an optimal decision threshold for a model in a systematic way that allows for fair comparison of other threshold dependent metrics. We elected to use the F-2 score (a weighted version of the F-score that gives preference to sensitivity), because a false negative (labeling a chronic patient as non-chronic) was considered more costly than a false positive. Models were also compared using the calibration intercept and slope, as calculated on the held-out testing set.

To explore the impact of potentially non-linear relationships in the RF model, we generated partial dependence plots (PDPs) for key predictors (20). PDPs are created using simulated patients to estimate the impact of changing a particular variable on the predicted risk generated by the model. This provides an approximation of the relationship between a given predictor and the outcome, and this estimate is non-parametric, allowing plots to depict the arbitrary, non-linear relationships characteristic of RF models. These plots do not account for potential correlations between variables, so they should be taken only as approximations of the actual relationships, which may be more complex.

In addition to direct comparison between models using the above metrics, we performed a net-benefit analysis (21) to evaluate whether the simplest (LR) or most complex (RF) models could improve decision-making at various levels of risk tolerance (Figure S4). This analysis compares the two models to two potential default treatment strategies: withholding opioids from all patients in order to avoid COU or providing opioids regardless of COU risk. The net-benefit of using the predictive model is calculated in terms of the percentage of the total patient population that would otherwise develop COU across a range of risk tolerances.

Transportability analysis

Because this dataset consisted of patient records from disparate HCOs, mostly in different US geographical regions, we were able to investigate the transportability of the model across organizations. Transportability refers to the tendency of a model to perform similarly when evaluated on a new sample of data which was not included in model training, such as performance in a new HCO or a later time period. Selecting a subset of 11 HCOs with a suitable sample size of patients (n >750), we performed a "leave-one-group-out" (LOGO) validation approach to estimate the performance of a model when implemented in an HCO that was not included in training. For each iteration of this analysis, the models were trained on 10 of the 11 HCOs, while the held-out HCO was used to evaluate performance. This process was repeated 11 times, holding out a different HCO each time. To determine whether the variance in model AUC across organizations was due to chance, we compared AUC of the same models trained and evaluated on 30 random data subsamples across all HCOs with the size of each sample adjusted to match the sizes of the HCO specific samples.

Results

Patient Characteristic

We identified 33,894 patients with OA and at least one opioid prescription during the study period who met inclusion criteria (Figure S1). Of these, 2,925 (8.6%) met the definition of COU during the first 12 months after the index prescription. These patients differed significantly from the non-chronic use patients across a wide array of characteristics (Table 1), including demographic, health-related, and socioeconomic factors. Among other differences, patients

with chronic use were younger, had higher rates of depression or anxiety, lower rates of joint replacement prior to index, and were more likely to be current smokers.

Model Performance

We developed and evaluated several models with increasing degrees of complexity (and decreasing interpretability) to evaluate the potential tradeoffs between model parsimony and discrimination. All models performed relatively well, with the AUC ranging from 0.72 to 0.76 across models for internal cross-validation in the training set (3 repeats of 10 folds, 2,237 patients per fold), or 0.70 to 0.73 for the held-out testing set (11,524 patients) (Table 2, Figure S2). The gap between the most complex (RF) and least complex (LR) models was significant but small, at 0.03 AUC using cross-validation or 0.01 in the testing set. We set decision thresholds for each model using F-2 score so that models could be compared directly (see methods). Using these thresholds, the models performed similarly in terms of sensitivity and specificity (Table 2).

We next evaluated calibration, which is the relationship between the predicted patient risk and the observed likelihood of the outcome, using the held-out testing dataset. We generated calibration plots to evaluate the proportion of patients with COU within different risk deciles. Predicted risk deciles spanned a clinically relevant range (10^{th} percentile to 90^{th} percentile for LR: 0.02–0.18; EN: 0.03–0.16; RF: 0.04–0.15; SVM: 0.18–0.68). We calculated the intercept and slope of a regression between predicted risk and the outcome of chronic use (where perfect calibration would reflect an intercept of 0 and a slope of 1) and E_{avg} which is the average absolute difference between a smooth calibration curve and the diagonal line of perfect calibration (Figure 2) (22). We found the LR (intercept 0.01, slope 0.81, E_{avg} 0.006) and

EN (intercept 0.00, slope 0.97, E_{avg} 0.006) models had good calibration while the RF model (intercept -0.04, slope 1.44, E_{avg} 0.016) showed fair calibration. The SVM model (intercept -0.05, slope 0.3, E_{avg} 0.346) had poor calibration, as this model was trained on rebalanced data (see Methods) which does not reflect the true incidence rate.

Predictor Importance

Based on the overall similarity in performance across models, we selected for further analysis the simplest (LR) and the most accurate (RF) models. We evaluated the most important predictors as determined by the odds ratio (LR) or impurity-based feature importance (RF) (19). Both models had a wide range of significant predictors, which spanned the complete set of predictor types (Figures 3A and 3B).

The most important predictors were substantially different between models, with only 6 of the top 20 predictors in common. To further explore these differences, we generated partial dependence plots (PDPs), which show the relationship between a given predictor and the outcome by marginalizing across all other predictors, for the most important features in the RF model (Figures 3C, 3D, and S3) (20). These plots use simulations to explore the relationships the RF model has learned between one of the predictors and the outcome, including if the predictor is non-linear (i.e., not monotonically increasing or decreasing). The PDPs showed several non-linear relationships. For instance, patients who were only very recently diagnosed (within the 14 days before index) with OA (according to first observed diagnostic code on a claim or problem list) (Figure 3C), and patients with <5 provider visits were predicted as having dramatically higher risk of chronic use (Figure 3D).

Model Transportability

We next investigated how models could be expected to perform in new datasets that were not part of the model development process (i.e., model transportability) using a LOGO validation approach (see methods). We found variability in model AUC across held out HCOs, spanning a range from 0.842 (excellent) to 0.571 (barely better than chance) (Figure 4A). The variance of model AUC across organizations was much higher than the expected variance based on models trained on identically sized random samples from the entire patient pool for both LR ($F = 7.52$, $p = 1.03$ e-5, two-sided F-test) and RF ($F = 7.48$, $p = 1.09$ e-5) models (Figure 4B). Although the variance across models was slightly higher for LR than RF in the LOGO analysis, this difference was not statistically significant ($F = 1.54$, $p = 0.50$).

Discussion

We developed and evaluated several types of predictive models using EHR data to estimate the risk of patients with OA developing COU within 1 year of a new opioid prescription. Multiple models offered meaningful predictive power, with the best performing RF model yielding an AUC of 0.73 on a held-out testing set. The LR and EN models had excellent calibration, i.e., the predicted risk closely approximated the true patient risk. Predictive models therefore offer a useful tool for clinical practice, where they could be used to estimate individualized risk of COU to better inform care of patients with OA.

A primary objective of this study was to determine whether complex black-box methods were better than simpler, more interpretable approaches. While the more complex RF model had slightly (though significantly) greater discrimination ability in the primary analysis, this difference was notably smaller than has been reported in some studies (9,23). While improved

discrimination can provide more accurate identification of at-risk patients, these benefits must be weighed against the ethical and practical concerns of implementing black-box models in the clinic (24). Given the relatively small difference in terms of model discrimination (especially in validation data), as well as better calibration for the LR model, we find that the more interpretable LR model is likely preferrable in this application.

While the current study was not designed to investigate causal relationships, some of the predictor relationships uncovered may be clinically relevant and suggest areas for future research. For instance, any recorded pain score, including severe pain, was associated with lower risk of COU. While the trend across pain levels was as expected (i.e., mild pain was associated with the lowest risk), failing to record the pain score in EHR (either at the individual level or HCO level) was associated with higher risk than any degree of documented patient-reported pain score, after accounting for other variables. Additionally, we found that patients with a very recent OA diagnosis or very few provider visits had dramatically higher predicted risk in the RF model (Figure 4). With the current dataset, we could not address whether this is indicative of patient behavior (e.g., switching providers to seek opioids) or provider behavior (e.g., assigning a diagnosis code only to patients with severe pain).

As a secondary analysis, we explored the performance of the LR and RF models when tested on HCOs that were not part of the training set using a leave-one-group out analysis. Model transportability, how well a model can be expected to perform in a new environment, is an increasingly important issue and is rarely investigated as part of normal model development for CPMs (25,26). The variability we observed across organizations suggests that even the relationships identified in a large, multi-center dataset should not be presumed to apply to

HCOs outside of that dataset. This is a critical consideration when implementing CPMs and is one reason why CPMs often have a smaller than expected impact on patient care (27,28). These results suggest that care should be taken when extending these (and other) CPMs into new health care settings which were not used as part of model development.

There are many clinical scenarios in which estimating the risk of COU in a subacute treatment setting may be beneficial for both the prescribing physician and the patient. An example of such a scenario may include a patient with severe osteoarthritis pain awaiting joint replacement surgery. A risk assessment, informed by the estimates produced by this model, could foster a discussion between the physician and patient about the risk that short term use of opioids may lead to COU. A net-benefit analysis (21) suggests that both the RF and LR models presented here can improve decision making across a range of risk tolerances (Figure S4). The difference between the two models is small, making the more interpretable LR model likely the best option for implementation in clinical practice. Ultimately, patients must be treated individually, and risk estimates such as those provided by predictive models can help providers and patients come to the best decision possible.

Limitations

As this study includes only patients with diagnosed OA who received at least one opioid prescription, the model may not be applicable to general OA patients, many of whom may not be formally diagnosed in the EHR, nor to other patient populations at risk for COU. Although efforts were made to include only patients receiving opioids to treat OA pain (e.g., by removing other common chronic pain conditions and surgical encounters), it is impossible to determine if patients received opioids for OA pain specifically. Finally, because this dataset is restricted to

EHR data from 13 HCOs, it is likely that the incidence of COU is underestimated if patients receive opioids from other sources (e.g., other HCOs) that are not represented in this dataset.

This study employed a unified database of EHR data from AMGA (American Medical Group Association) member organizations as mapped by Optum®, and this mapping process imposed uniformity on the data structure across HCOs. This is relevant for the discussion of transportability, as HCOs within this environment are likely to be more similar to one another than they are to outside HCOs. Therefore, the variability across organizations may have been even larger if the analysis was extended to additional HCOs.

Acknowledgements

We would like to thank Barbara Kaplan Pritchard and Jaimee Reiley for their support with the development of this project, Birol Emir and Xiang (Jay) Ji for statistical support, Caitlin Shaw for assistance with data extraction and technical support, and Jannette Escobar and Cindy Shekailo for project management and administrative support.

References

1. Cisternas MG, Murphy L, Sacks JJ, Solomon DH, Pasta DJ, Helmick CG. Alternative Methods for Defining Osteoarthritis and the Impact on Estimating Prevalence in a US Population-Based Survey. *Arthritis Care Res* 2016;68:574–580.

2. Anon. The Burden of Musculoskeletal Diseases in the United States, Fourth Edition. Available at: https://www.boneandjointburden.org/fourth-edition/iiib10/osteoarthritis. Accessed August 26, 2021.

3. Thorlund JB, Turkiewicz A, Prieto-Alhambra D, Englund M. Opioid use in knee or hip osteoarthritis: a region-wide population-based cohort study. *Osteoarthr Cartil* 2019;27:871–877.

4. Rudd RA, Seth P, David F, Scholl L. Increases in Drug and Opioid-Involved Overdose Deaths — United States, 2010–2015. *MMWR Morb Mortal Wkly Rep* 2016;65.

5. Miller M, Stürmer T, Azrael D, Levin R, Solomon DH. Opioid analgesics and the risk of fractures in older adults with arthritis. *J Am Geriatr Soc* 2011;59.

6. O'Neil CK, Hanlon JT, Marcum ZA. Adverse effects of analgesics commonly used by older adults with osteoarthritis: Focus on non-opioid and opioid analgesics. *Am J Geriatr Pharmacother* 2012;10:331–342.

7. Krebs EE, Gravely A, Nugent S, Jensen AC, DeRonne B, Goldsmith ES, et al. Effect of opioid vs nonopioid medications on pain-related function in patients with chronic back pain or hip or knee osteoarthritis pain the SPACE randomized clinical trial. *JAMA - J Am Med Assoc* 2018;319:872–882.

8. Kolasinski SL, Neogi T, Hochberg MC, Oatis C, Guyatt G, Block J, et al. 2019 American College of Rheumatology/Arthritis Foundation Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee. *Arthritis Rheumatol* 2020;72:220–233.

9. Dong X, Deng J, Hou W, Rashidian S, Rosenthal RN, Saltz M, et al. Predicting opioid overdose risk of patients with opioid prescriptions using electronic health records based on temporal deep learning. *J Biomed Inform* 2021;116:103725.

10. Dunn KM, Saunders KW, Rutter CM, Banta-Green CJ, Merrill JO, Sullivan MD, et al. Overdose and prescribed opioids: Associations among chronic non-cancer pain patients. *Ann Intern Med* 2010;152:85–92.

11. Muntner P, Colantonio LD, Cushman M, Goff DC, Howard G, Howard VJ, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *JAMA - J Am Med Assoc* 2014;311.

12. Unger T, Borghi C, Charchar F, Khan NA, Poulter NR, Prabhakaran D, et al. 2020 International Society of Hypertension Global Hypertension Practice Guidelines. *Hypertension* 2020;75:1334–1357.

13. Anon. Introduction: Standards of Medical Care in Diabetes—2019. *Diabetes Care* 2019;42:S1 LP-S2.

14. Subedi M, Bajaj S, Kumar MS, YC M. An overview of tramadol and its usage in pain management and future perspective. *Biomed Pharmacother* 2019;111:443–451.

15. Karmali RN, Bush C, Raman SR, Campbell CI, Skinner AC, Roberts AW. Long-term opioid therapy definitions and predictors: A systematic review. *Pharmacoepidemiol Drug Saf* 2020;29:252–269.

16. Dunn KM, Saunders KW, Rutter CM, Banta-Green CJ, Merrill JO, Sullivan MD, et al. Opioid prescriptions for chronic pain and overdose: A cohort study. *Ann Intern Med* 2010;152:85–92.

17. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019;170.

18. Boonstra AM, Stewart RE, Köke AJA, Oosterwijk RFA, Swaan JL, Schreurs KMG, et al. Cut-Off Points for Mild, Moderate, and Severe Pain on the Numeric Rating Scale for Pain in Patients with Chronic Musculoskeletal Pain: Variability and Influence of Sex and Catastrophizing. *Front Psychol* 2016;0:1466.

19. Kuhn M, Johnson K. *Applied predictive modeling*.; 2013.

20. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001;29.

21. Vickers AJ, Calster B Van, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352.

22. Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer-Verlag; 2001.

23. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. In: *Critical Care Medicine*.Vol 44. Lippincott Williams and Wilkins; 2016:368–374.

24. Ozaydin B, Berner ES, Cimino JJ. Appropriate use of machine learning in healthcare. *Intell Med* 2021;5:100041.

25. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat Commun* 2020;11:1–12.

26. Pajouheshnia R, Smeden M van, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol* 2019;105:136–141.

27. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA* 2018;319:1317–1318.

28. Upshaw JN, Nelson J, Koethe B, Park JG, McGinnes H, Wessler BS, et al. Performance of Heart Failure Clinical Prediction Models: A Systematic External Validation Study. *medRxiv* 2021:2021.01.31.21250875.

Table 1: *Select baseline population characteristics stratified by incident chronic opioid use*

Characteristics	*Non-chronic Opioid Use*	*Chronic Opioid Use*
n	**30969**	**2925**
Age, mean (SD), years	62.4 (11.8)	61.4 (11.8)
Male, n (%)	11782 (38.0)	1134 (38.8)
Race, n (%)		
White or caucasian	26903 (86.9)	2451 (83.8)
Black or african american	1957 (6.3)	217 (7.4)
Other	592 (1.9)	58 (2.0)
Unknown race	1517 (4.9)	199 (6.8)
Ethnicity, n (%)		
Not hispanic, latino or spanish origin	27125 (87.6)	2556 (87.4)
Hispanic, latino or spanish origin	1109 (3.6)	91 (3.1)
Unknown ethnicity	2735 (8.8)	278 (9.5)
Smoking status, n (%)		
Previously smoked	9111 (29.4)	898 (30.7)
Current smoker	4486 (14.5)	695 (23.8)
Never smoked	14210 (45.9)	1030 (35.2)
Unknown	1772 (5.7)	195 (6.7)
Not currently smoking	1390 (4.5)	107 (3.7)
Insurance Type, n (%)		
Commercial	12415 (40.1)	960 (32.8)
Medicare	14922 (48.2)	1501 (51.3)
Other non-govt	1967 (6.4)	159 (5.4)
Medicaid	1288 (4.2)	273 (9.3)
Other govt.	377 (1.2)	32 (1.1)
Median income in ZIP code, mean (SD)	59133.9 (19570.1)	56125.2 (18165.4)
Number of PCP office visits, mean (SD)	4.1 (4.8)	3.6 (5.1)
Number of ER visits, mean (SD)	0.6 (1.6)	0.5 (1.6)
Depression Dx, n (%)	6958 (22.5)	889 (30.4)
Anxiety Dx, n (%)	6853 (22.1)	890 (30.4)
Categorized pain score,[1] n (%)		
mild	6677 (21.6)	483 (16.5)
moderate	3614 (11.7)	368 (12.6)
severe	3849 (12.4)	418 (14.3)
missing for health care organization	5575 (18.0)	553 (18.9)
missing for individual	11254 (36.3)	1103 (37.7)
Hip OA Dx[2], n (%)	3691 (11.9)	394 (13.5)
Knee OA Dx[2], n (%)	13289 (42.9)	1189 (40.6)
Shoulder OA Dx[2], n (%)	3707 (12.0)	315 (10.8)
Other OA Dx[2], n (%)	12646 (40.8)	1261 (43.1)
Polyarthritis Dx, n (%)	3430 (11.1)	474 (16.2)
Joint replacement Px, n (%)	5212 (16.8)	203 (6.9)

Dx: Diagnosis; Px: Procedure; PCP: Primary care provider; ER: Emergency room

[1]Numeric pain score was categorized into levels mild (0-5), moderate (6-7), and severe (8-10)

[2]Conditions are non-exclusive, so percentages will not sum to 1

Table 2: *Prediction performance of all models*

	Held-out Validation	Internal Cross-Validation			
Prediction Model	AUC	AUC	Tuned Threshold	Sensitivity	Specificity
Random Forest	**0.728 (0.711-0.745)**	**0.756 (0.746-0.765)**	0.102	**0.648 (0.638-0.659)**	**0.731 (0.726-0.736)**
Elastic Net	0.717 (0.699-0.736)	0.731 (0.721-0.740)	0.094	0.632 (0.622-0.643)	0.721 (0.717-0.725)
Logistic Regression	0.717 (0.699-0.734)	0.729 (0.720-0.738)	0.089	0.645 (0.635-0.654)	0.703 (0.699-0.706)
Support Vector Machine	0.702 (0.683-0.720)	0.723 (0.714-0.732)	0.523	0.668 (0.657-0.679)	0.661 (0.657-0.666)

AUC: Area under the receiver operating characteristic curve; (95% confidence interval)

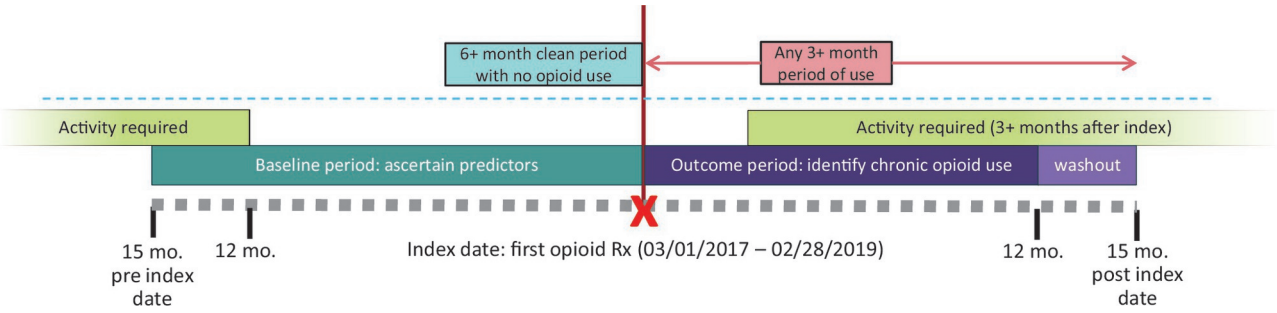
Figure Legends

Figure 1: *Study Schema.* Patients were required to have evidence of activity in the EHR ≥1 year prior to the index date, and at least 3 months later than the index date. Patients were also required to have at least a 6-month clean period, indicating that no opioid prescriptions were recorded immediately prior to the index date. Baseline characteristics were ascertained using data from the 15 months prior to the index date. The period of chronic use could be initiated on the index date or at any point within the one-year outcome period. A 3-month washout period was included to capture prescriptions relevant for chronic use periods initiated near the end of the 12-month outcome period.

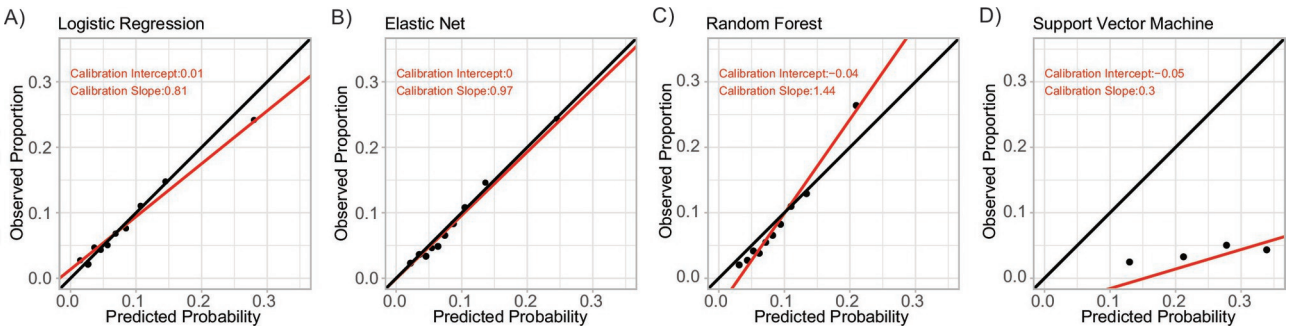
Figure 2: *Calibration plots for LR and RF models.* Patients are split into deciles according to predicted risk. The average predicted risk within in each decile is then compared to the actual incidence. Calibration intercept and slope are shown in red, fit by regressing the outcome variable against the predicted log odds across deciles. The black unity line indicates perfect calibration.

Figure 3: *Predictor importance* (A) The 20 most impactful patient features (by coefficient magnitude) are shown for the LR model. (B) The 20 patient features with the highest Gini-based importance are shown. For both plots, colors group patient features into categories. (C and D) Partial dependence plots provide an estimate of the relationship between duration of OA (C) or number of EM visits (D) and the risk of COU. Below, population density plots show the distribution of the study population. EM: evaluation and management visit; ED: emergency department; IP: inpatient; DX: diagnosis; RX: prescription; SDOH: social determinants of health; NSAID: non-steroidal anti-inflammatory drug; APAP: acetaminophen.

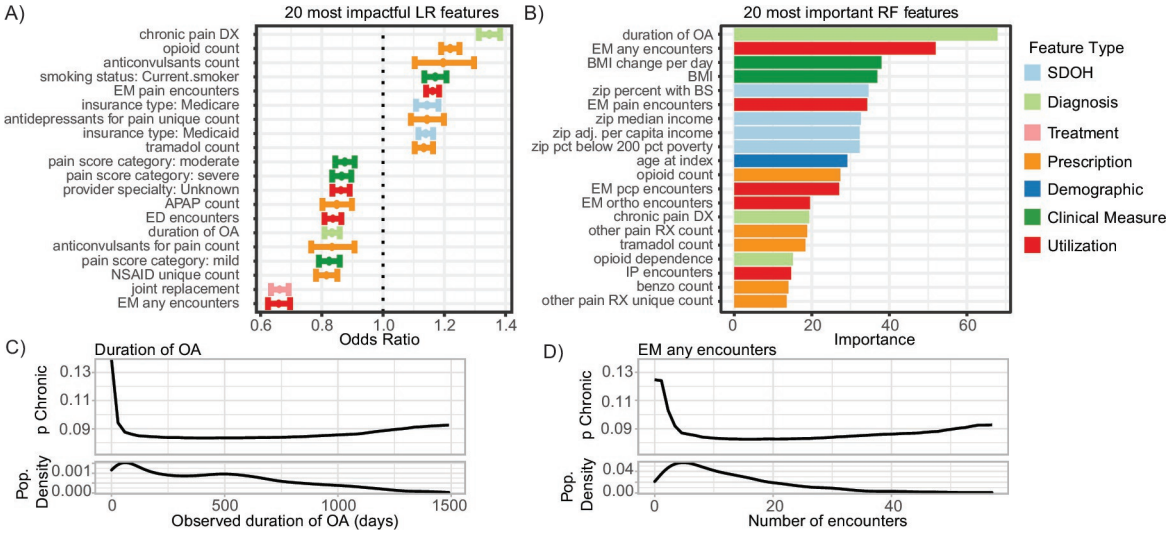
Figure 4: *Transportability of models across HCOs* (A) The performance of each model was evaluated using a leave-one-group-out approach to estimate the discrimination ability of the model when implemented in a health system outside the training set. Each iteration involved training on all but one HCO, and then calculating AUC on the remaining HCO. (B) variability in performance across random samples which were not split based on organization (i.e., all HCOs are included in both training and testing sets).



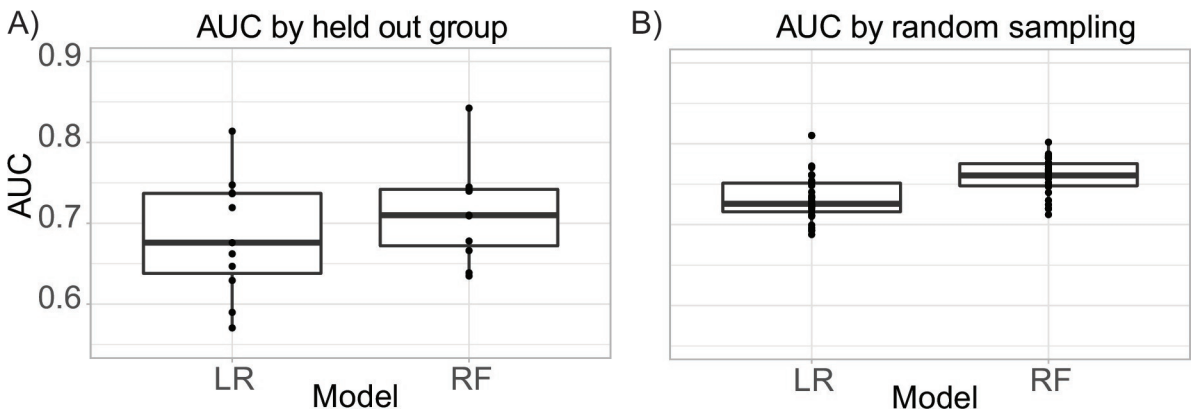
acr_25013_f1_study_schema.eps



acr_25013_f2_calibration.eps



acr_25013_f3_variables_pdf_resize.eps



acr_25013_f4_logo_comparison.eps