Journal of Business Analytics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjba20

Short-term prediction of opioid prescribing patterns for orthopaedic surgical procedures: a machine learning framework

Ebrahim Mortaz, Ali Dag, Lorraine Hutzler, Christopher Gharibo, Lisa Anzisi & Joseph Bosco

To cite this article: Ebrahim Mortaz, Ali Dag, Lorraine Hutzler, Christopher Gharibo, Lisa Anzisi & Joseph Bosco (2021) Short-term prediction of opioid prescribing patterns for orthopaedic surgical procedures: a machine learning framework, Journal of Business Analytics, 4:1, 1-13, DOI: 10.1080/2573234X.2021.1873078

To link to this article: https://doi.org/10.1080/2573234X.2021.1873078

Published online: 17 Jan 2021.

Submit your article to this journal

Article views: 176

View related articles

View Crossmark data

JOURNAL OF BUSINESS ANALYTICS
2021, VOL. 4, NO. 1, 1–13
https://doi.org/10.1080/2573234X.2021.1873078

ORIGINAL ARTICLE

Short-term prediction of opioid prescribing patterns for orthopaedic surgical procedures: a machine learning framework

Ebrahim Mortaz[a], Ali Dag[b], Lorraine Hutzler[c], Christopher Gharibo[d], Lisa Anzisi[e] and Joseph Bosco[f]

[a]Department of Management Science, Pace University, New York, NY, USA; [b]Heider College of Business, Creighton University, Omaha, NE, USA; [c]Quality, of Orthopedic Surgery, NYU Langone Health, New York, NY, USA; [d]Department of Anesthesiology, Perioperative Care, Orthopedics, NYU Langone Health New York, New York, NY, USA; [e]Pharmacy Utilization Management NYUPN Clinically Integrated Network, New York, NY, USA; [f]Orthopedics, NYU Langone Orthopedic Hospital, NYU Langone Health, New York, NY, USA

ABSTRACT
Overprescribing of opioids after surgical procedures can increase the risk of addiction in patients, and under prescribing can lead to poor quality of care. In this study, we propose a machine learning-based predictive framework to identify the varying effects of factors that are related to the opioid prescription amount after orthopaedic surgery. To predict the prescription classes, we train multiple classifiers combined with random and SMOTE oversampling and weight-balancing techniques to cope with the imbalance state of the dataset. Our results show that the gradient boosting machines (XGB) with SMOTE achieve the highest classification accuracy. Our proposed analytical framework can be employed to assist and therefore, enable the surgeons to determine the timely changing effects of these variables.

ARTICLE HISTORY
Received 28 July 2020
Accepted 26 December 2020

KEYWORDS
Opioid prescription; machine learning; feature extraction

1. Introduction

Drug overdose is the leading cause of accidental death in the US, with 52,404 lethal drug overdoses in 2015 (Rudd, 2016). Opioid addiction drives this epidemic, with 12,990 related to heroin overdose and 20,101 related to overdose due to prescription opioids (Rudd, 2016). Pain management is an important aspect of patient care and opioids are often prescribed for pain relief following most surgical procedures. Orthopaedic patients may experience substantial pain after surgical procedures, hence orthopaedic surgeons are the third highest prescribers of opioids among all physicians (Morris et al., 2015). However, misuse or abuse of opioid by patients can bring about many unintended consequences.

Numerous challenges exist when it comes to implementation of postoperative opioid protocols. First, the level of discomfort a patient experiences after surgery varies between individuals. Second, the opioid dosing and duration vary between physicians and third, there is a lack of consensus among the experts about the number of days after which the risk of addiction to opioids increases (National Academies of Sciences, Engineering and Medicine, 2017). In 2016, the CDC proposed a protocol defining certain guidelines for opioid prescription days and doses for chronic pain (Dowell et al., 2016). The guidelines provide recommendations for primary care clinicians who are prescribing opioids for chronic pain which is outside of cancer treatment, palliative care, and end-of-life care. Based on the CDC guidelines, for the number of prescription days, the lowest effective dosage for 3 days or less is often sufficient. For daily prescription doses, given in morphine milligram equivalent (MME) dosing, the guidelines include four ranges of MMEs. The ranges are < 50 MME, 50–90 MME, 90–120 MME and > 120 MME. For anyone > 50 MME/day, the clinicians are advised to offer naloxone. All dosages and especially the higher dosages require careful justification based on benefit and risks.

As noted, the CDC guidelines are primarily intended for chronic pain management. For acute postoperative pain, the CDC refers the clinicians to the instructions provided by the where it advises to reserve the use of opioids for moderate to severe acute pain and prescribe the lowest possible doses if must be used. However, these blanket recommendations encompass all the patients with moderate to severe pain and do not offer any analytical method to understand and address the needs of each individual patient. Hence, there is an imperative need for an opioid prescribing system that personalises the patient's care. This system should steer clear of both under- and over-prescription and should automate the prescription process. Overprescribing can increase the risk of misuse and under prescribing can lead to poor quality of care which could also prevent a patient from participating in physical therapy on the way to full recovery.

The ML algorithms can recognise the underlying patterns in the data, and due to their iterative learning process can easily adapt as the new data emerges

CONTACT Ali Dag alidag@creighton.edu Heider College of Business, Creighton University, Omaha, NE, USA

(Bishop, 2006). Our predictive model will allow the surgeon to see the collective opinion of all physicians whose data was used to train the algorithms. The model will operate as a virtual agent which will automate and personalise the prescribing process. To keep the model practical with enough flexibility for the end-users, we treat the prediction process as a classification and not a regression model. That is, although the prescribed dose of opioids in MMEs/day units are available, we choose to predict the prescription classes and not the real-valued doses. We believe predicting real-valued doses could diminish the attractiveness of the proposition for adoption by practitioners. These prescription classes are obtained from the thresholds provided by the CDC guidelines. Although the CDC guidelines have shown multiple implementation challenges, might be re-evaluated by the agency in future (Kroenke et al., 2019), and are intended for chronic pain management, we adopt the offered thresholds to be able to classify higher MMEs doses and identify the underlying patterns within the patients' cohort. We should also note that if the current-offered thresholds by the CDC change in the future, the proposed predictive framework could easily be adapted.

Our focus is on orthopaedics procedures which include knee, shoulder, hip and revision arthroplasty surgeries. After training multiple algorithms and finding the best model, the next goal would be to deploy the proposed framework as an open access application, such as a web app or a mobile app, for dissemination among healthcare providers. In addition, identifying the most important features that have the highest impact on prediction accuracy is the other goal of this study. The feature selection will allow the surgeons to parse the outcomes in prescribing a certain dose after surgery and assist the policymakers understand the important features when developing prescription protocols.

Our research distinguishes from the existing opioid prescription studies from two main aspects. We aim to propose a machine learning-based predictive framework, which.

- determines the dosage of the opioid to prescribe according to the MME guidelines for Day 1, 2, 3 and 4.
- identifies the varying (dynamic) effects of factors on opioid usage that are related to opioid prescription amount (1, 2, 3 and 4 days) after orthopaedic surgery.

To the best of our knowledge, the dynamic effects of the opioid usage predictors has never been studied in the literature, which would potentially enable the care givers to prescribe daily opioid amounts in a customised manner instead of choosing the lowest boundary amount of a wide range that is dictated by the current postoperative opioid protocols. Thus, our study not only takes a wide set of variables into account independent of the time, but it also analyzes their affects in a temporal manner (from day 1 to 4) and differentiates whose effect (on opioid use) changes over 4 days after the surgery.

2. Literature review

The opioid abuse crisis has been widely studied from different perspectives for different surgical and non-surgical practices. As we aim to predict opioid prescription classes using ML classifiers, we divide the literature review section into two streams; a) *Opioid problem*, and b) *Machine learning methodology*. The *opioid problem* stream studies the papers that have addressed the opioid problem but not necessarily used ML models. The *machine learning methodology* stream studies the papers that have used ML models on different but similar healthcare problems.

2.1. Opioid problem

We mainly focus on the papers that discuss the orthopaedic surgical procedures. Lo-Ciganic et al. (2019) presented a more general perspective where opioid overdose risk among Medicare beneficiaries is predicted. The majority of the papers that are reviewed in this stream have studied the opioid problem from risk of misuse/abuse perspective.

Schoenfeld et al. (2018) developed one of the few ML-based frameworks to predict the risk of opioid abuse after spine surgery. The goal was to identify at-risk/not of abuse patients in the long term (years). The gradient boosting algorithm offered the best accuracy performance in their analysis. It was reported that the preoperative opioid duration, antidepressant use, tobacco use, and Medicaid insurance are the most important predictors of sustained postoperative opioid prescriptions after spine surgery. The remainder of the reviewed publications has deployed statistical analysis.

Rhon (2018) conducted a multivariate logistic regression for patients who underwent arthroscopic hip procedure to identify variables that predicted chronic opioid use 2 years after surgery. Socioeconomic status, prior use of opioid medication, prior use of non-opioid pain medication, high health-seeking behaviour before surgery, insomnia and mental health disorder were all predictive of chronic opioid use.

Cook et al. (2017), developed a cluster analysis to group pre-operative (12 months) and post-operative (24 months) patients' population on opioid prescription patterns in non-arthroplasty orthopaedic hip surgery. The clustering was based on Bayesian information criterion which provided two distinct clusters for short-duration, high supply and long-duration, lesser supply.

Multiple studies have assessed the predicting factors of chronic opioid use in arthroplasty procedures. Goesling et al. (2016) studied the trends in persistent opioid use risk after total knee and total hip arthroplasty surgeries and concluded that persistent opioid use was not associated with change in joint pain for opioid naïve and opioid-dependent patients. The analysis was based on statistical testing.

Kee et al. (2016) similarly reviewed the overall risk factors for opioid use in chronic pain patients who underwent orthopaedic surgeries. But, unlike the previous study, the authors provided only general recommendations for the orthopaedic surgeon's role in managing complicated patients.

Kim et al. (2017) studied the patterns and predictors of persistent opioid use after arthroplasty procedures using multivariate statistical analysis. It was concluded that in the knee arthroplasty vs hip, a longer hospitalisation, discharge to a rehabilitation facility, preoperative opioid use, higher comorbidity, and back pain have the highest impact for persistent opioid use.

Sing et al. (2016), investigated the impact of preoperative opioid use on early outcomes in total joint arthroplasty. It was shown that preoperative opioid use should be disclosed as a risk factor and taken into consideration by physicians before initiating pain management by opioids.

In a similar studies conducted by Zywiel et al. (2011) and Zarling et al. (2016), the impact of preoperative opioid use on the outcomes of the surgery and persistent opioid use after surgery is examined. They collectively indicated that the preoperative opioid use should be disclosed as a risk factor before initiating opioid management as the patients who chronically use opioid medications prior to arthroplasty surgeries are at a greater risk for complications and longer recovery periods.

A summary of the papers we discussed is presented in Table 1 to better differentiate the contributions presented in our paper. In short, none of the reviewed papers provide a quantitative recommendation system for opioid prescription (how much to prescribe). They primarily studied the risk of abuse/overdose/chronic use and what predicters impact that risk in long term. What we offer is a prediction system to determine how much to prescribe following selected orthopaedic surgeries.

2.2. Machine-learning methodology

The ML algorithms work markedly on the data with abundant features and complicated patterns and therefore have extensively been used in health analytics problems. Transplantation of organs such as heart (Dag et al., 2017, 2016; Dolatsara et al., 2020), lung (Oztekin et al., 2018) and kidneys (Kusiak et al., 2005; Topuz et al., 2018), or studying readmission probabilities after procedures (Nasir et al., 2019; Zheng et al., 2015), hospital no show probabilities (Nasir et al., 2020, Simsek, Tiahrt, et al., 2020), survivability probabilities after procedures (Dursun et al., 2005), surveillance for opioid misuse (Prieto et al., 2020) and cancer survival probabilities (Simsek, Kursuncu, et al., 2020) are all among various healthcare problems that have taken advantages of ML-based methodology. The number of publications in this area is very large and for brevity, we only review the papers that have taken steps akin to our predictive framework. Dag et al. (2016) proposed a Bayesian believe network for scoring the preoperative recipient-donor heart transplant survival. They were able to study the interactions among the predictors and provide individualised survival score. Again, Dag et al proposed a ML-based framework for predicting heart transplantation outcomes in 2018. The goal was to predict the transplantation survival for different stages after the surgical procedure (1, 5 and 9 years). The impact of features over several stages was studied by a grouping mechanism. Dolatsara et al. (2020) also studied the heart transplantation survival problem with a ML-based methodology. But, compared to the Dag et al's study, the authors additionally estimated the survival probabilities over time using isotonic regression. For kidney graft survival, Topuz et al. (2018) used ML models and Bayesian believe network, similar to the study by Dag et al, to predict the graft survivability and study the conditional probabilities among predictors. For patients undergoing lung transplants, Oztekin et al. (2018) devised a data analytic framework based on ML models to predict the quality of life. This study

Table 1. Literature review on the opioid problem in orthopaedics procedures.

Study	Objective of the study	Procedure	Methodology
Lo-Ciganic et.al.	Risk of overdose	General	Machine learning
Schoenfeld et.al.	Risk of abuse	Spine	Machine learning
Rhon et.al	Risk of chronic use	Total hip	Statistical testing
Cook et. al.	Group similar patients	Total hip	Cluster analysis
Goesling et. al.	Identify risk factors	Total knee and hip	Statistical testing
Kee et. al.	General/chronic pain	General	Statistical testing
Kim et. al.	Risk of chronic use	Total hip or knee	Multivariate analysis
Sing et. al.	Risk of abuse	Total joint	Statistical testing
Zarling at. al	Risk of abuse	Total joint	Statistical testing
Zywiel et. al.	Risk of chronic use	Total knee	Statistical testing
Our study	Prescription recommender	Hip, knee, shoulder and revisions	Machine learning

took an additional step and applied genetic algorithms for feature reduction in the dataset. The goal of the feature reduction was to reduce the number of features to accelerate the training process. Also, related to *opioid problem* stream, the study by Schoenfeld et al. (2018) used ML models for opioid dependency prediction after spine surgery. The goal of the study was to identify the at-risk patients perioperatively. Recently in 2020, Prieto et al proposed an ML-based framework to enhance the effectiveness of finding opioid misuse and heroin overdose among paramedic field responses. The offered framework was to improve responses for prevention of overdoses and opioid-related problems.

3. Methodology

The proposed machine-learning framework in this study consists of three phases in sequence which are shown in Figure 1. The three phases are 1) preparing the data for analysis, 2) training classification algorithms, and 3) feature selection.

In Phase I, the data is cleaned, that is the missing values are checked in, the calculated morphine milligram equivalent (MMEs) are classified into MME ranges, and the categorical variables are encoded into numerical ones. We log-transformed skewed features, grouped neighbour zip codes (6 groups overall) and engineer new features (surg-hist and median income). In addition, because the dataset is imbalance, we take advantage of two resampling techniques, random over sampling (ROS) and Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). We choose the two sampling methods due to their satisfactory initial results.

Phase II conducts the classification experiments. The classification algorithms in Phase II that are considered to experiment are K-nearest neighbour (KNN), ridge regression (RR), logistic regression (LR), support vector machines (SVM), random forest (RF), gradient boosting machine (GBM), and neural networks (NN). The considered classification algorithms belong to various classes of statistical, analytical and deep learning methods. All the trained algorithms are validated through 10 repeats of 10-fold cross validation and evaluated by five accuracy metrics. The algorithms that have lower than a threshold initial accuracy are discarded and the remaining ones advance to feature selection phase. We choose to have threshold check at the end of this phase because of the time-consuming process of solving large optimisation problems.

Finally, in Phase III, the feature selection process based on genetic algorithm and node impurity metrics are implemented. The feature selection removes the noisy/irrelevant features if exist and improve the quality of fit. Fewer number of features is always desirable as it makes the model simpler and improves its interpretability.

3.1. Dataset

The dataset was acquired from a large single speciality orthopaedics hospital. The data contained 10,520 records with 9436 unique patient identifiers (*mrns*) and 22 features from year 2016. All the prescribed doses of opioids from day 0 to day 4 are given (MME/day). The *day 0* data included mixed IV and oral prescriptions thus we excluded this variable from our study. The *MME D1* to *MME D4* variables, MMEs for day 1–4, are assumed to be the target variables and the remainder are independent features. The information for the features is given in Table 2. Note that the income feature is an estimate for the patient's *income* based on demographics data, and therefore there includes an *income error* feature.

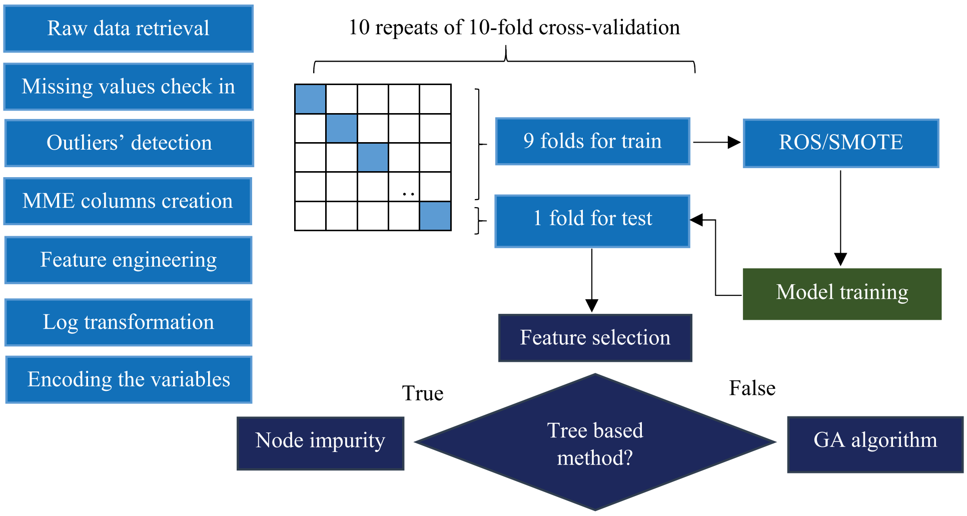


Figure 1. High level presentation of the predicting methodology.

Table 2. The characteristics of the dataset.

Variable	count (unique)/count (mean, IQR)
Medical record number (mrn)	10520 (9436)
sex	10520 (3)
procedure group (procedure)	10517 (11)
extended procedure group (e-proc)	10517 (16)
day0 mme (MME D0)	9943 (47.66, 45 – 15)
day1 mme (MME D1)	9210 (55.65, 67.50–22.50)
day2 mme (MME D2)	6690 (45.04, 52.50–2.50)
day3 mme (MME D3)	2983 (43.51, 52.50–2.50)
day4 mme (MME D4)	1257 (48.10, 60.00–2.50)
Length of stay (los)	10520 (2.40, 3–0)
age	10520 (64.65, 72–58)
BMI	10217 (34.59–15.20)
asa physical status (asa)	10499 (2.43, 3–1)
smoking status (smoking)	10520 (12)
race	10516 (7)
religion group (religion)	10515 (17)
zipcode	10517 (951)
surgeon	10520 (56)
discharge disposition (dch-disp)	10355 (10)
household income 2016 (income)	10464 ($75k, 51 USDk – 98 USDk)
margin of error of household income (income err)	10456 ($5k, 1k – 6k)
Insurance type (insurance)	10520 (8)
surgical history (surg-hist)	10519 (0.05, 0–0) [max = 2]

3.2. Phase I – data preparation (blue coloured)

To prepare the data for model training, we run four stages of cleaning, transformation, encoding and resampling. For cleaning, we remove the patients with missing values under *day 0* and patients with zero *los*. For transformation, we log-transform the numerical variables, *los, age, BMI*, and *income* to adjust skewness. For encoding, we classify *MME D1* to *MME D4* into five MME ranges following the protocol offered by the CDC. Note that we broke down 0–50 MMEs/day range into 0–30 MMEs/day and 30–50/day MMEs to better represent the low doses in short term predictions for the procedures. The frequencies of ranges within each day are given in Table 3.

For validation, we use 10 repeats of 10-fold cross validation. In k-fold cross validation introduced in (Kohavi, 1995), the entire dataset is randomly split into *k* mutually exclusive samples of approximately equal size and the prediction is tested *k* times using the *k* test sets. We also ensured the samples are stratified. The performance measure of the model is the average of the obtained performance measures. The 10 repeats will improve the stability of the models (Kuhn & Johnson, 2013).

Table 3. The prediction target ranges/classes for each day (MMEs/day).

day	(0–30)	(30–50)	(50–90)	(90–120)	(120+)	Total
1	2322	2531	2319	538	816	8826
2	2636	1809	1267	299	333	6344
3	1262	693	505	128	137	2725
4	477	290	215	61	79	1122

3.3. Phase II – classification models (green coloured)

The prepared data in Phase I is used to train seven ML algorithms for experimentation. The accuracy of the algorithms on the test sets are obtained and compared. The algorithms with poor performance are removed and the remaining ones advance to Phase III of feature selection.

3.4. Phase III – feature selection (dark coloured)

Identifying the most important features is to simplify the models and improve the interpretability and applicability of the proposition. Most of tree-based learning algorithms, such as random forests, offer a mechanism to identify/rank the important features, but that is not the case for all classifiers. For the classifiers that pass the accuracy threshold test in Phase II, but do not have feature selection capabilities, we use the genetic algorithm (GA). The GA method is a popular metaheuristic that solves optimisation problems that are too difficult or impossible to be solved with exact methods and off-the-shelf software. The GA usually gives near optimal solutions in a reasonable computational time. It has been used for features selection before in different problems (Leardi et al., 1992; Oztekin et al., 2018). The optimisation problem that is usually optimised in feature selection is a combination of a number of involved features and the accuracy of the performing model. Table 4 shows different steps of the GA. Note that a chromosome is a one-dimensional array with the number of elements equal to the number of features in the problem. The elements of the chromosome are binary values (0/1). If a feature is

Table 4. The GA steps.

```
rep = 0
    While rep < max_rep
    Split the data into k-stratified folds
    For each fold k:
    gen = 0
    Initialise a random population of chromosomes
    Train the learning algorithm using the train set
    Compute the fitness function for each chromosome using the test set
    Save the best chromosome
    While gen < max_gen
    Select $\Omega$ number of parents for the next step
    Apply crossover and mutation
    Replace population by parents and $\beta$ children
    Update the best chromosomes set
    gen+ = 1
    Break the while loop
    Save and store the best chromosome
    k+ = 1
    End for
    rep+ = 1
    Break the while loop
    Return the best chromosome set (max_rep × k stored best
    chromosomes)
```

present in the chromosome, the corresponding element is 1, and 0 otherwise.

In the given steps in Table 4, *rep* denotes the repeat number (0, … max_rep) of cross validation, population stands for a collection of chromosome (two dimensional array), fitness function is given in (1), the best chromosome is the one with the highest fitness function value, and *gen* denotes the generation number in the genetic algorithm (0, … max_iter). The two functions, crossover and mutations are activated in generating the children from the parents. In crossover, the two parents' gens (binary elements of the chromosome) are swapped from a random point in parents' chromosomes and a child is created. In mutation, random genes are inverted in the child's chromosome to diversify the search space. The genetic algorithms steps in detail have also been described in (Oztekin et al., 2018).

The fitness function of the GA that we considered in this study is:

max f = α (measure of accuracy) + (1-α) (no. of features in chromosome/total no. of features) (1)

α is the classification weight that determines the importance of accuracy measure against the number of features involved in the classification model. The higher the α, the more important the accuracy measure becomes.

3.5. Accuracy metrics

There has been an extensive debate on how to accurately measure and interpret the quality of fit for a classifier algorithm on a multi-class imbalance dataset. It is shown that for imbalance datasets, accuracy alone can be misleading and thereby multiple measures should be considered to properly evaluate the performance. Precision, recall, F score, (Shreve et al., 2011; Sokolova & Lapalme, 2009) and class balance accuracy (CBA) (Mosley, 2013) are all among the popular metrics to evaluate the performance of algorithms in imbalanced datasets. We also use imbalance accuracy, which is built up on the previous metrics, and introduced in (Mortaz, 2020) for evaluation. The formula for each metric is given in (2)-(8).

Let us assume C^k is a k by k confusion matrix of a classifier and c_{ij} is the element in row i and column j of C^k where $i, j = 1, 2, \ldots, k$. Let us also define $\mathrm{c}_{\mathrm{i}.}$, and $\mathrm{c}_{.\mathrm{i}}$ as below:

$$c_{i.} = \sum_{j=1}^{k} c_{ij} \tag{2}$$

$$c_{.i} = \sum_{i=1}^{k} c_{ij} \tag{3}$$

Then, we can define macro average precision, macro average recall, macro average F score, CBA and imbalance accuracy as below:

Overall accuracy:

$$\mathrm{ACC} = \frac{\sum_{\mathrm{i}} c_{ii}}{\sum_{\mathrm{i,j}} c_{ij}} \tag{4}$$

Macro average precision:

$$\mathrm{MAP} = 1/\mathrm{k}\sum_{i=1}^{k} \frac{c_{ii}}{c_{.i}} \tag{5}$$

Macro average recall:

$$\mathrm{MAR} = 1/\mathrm{k}\sum_{i=1}^{k} \frac{c_{ii}}{c_{i.}} \tag{6}$$

Class balance accuracy

$$\mathrm{CBA} = 1/\mathrm{k}\sum_{i}^{k} \frac{c_{ii}}{\max(c_{.i}, c_{i.})} \tag{7}$$

Imbalance accuracy metric

$$\mathrm{IAM} = 1/\mathrm{k}\sum_{\mathrm{i}}^{\mathrm{k}} \frac{\min\left(c_{ii} - \sum_{j \neq i}^{k} c_{ij}, c_{ii} - \sum_{i \neq j}^{k} c_{ij}\right)}{\max(c_{.i}, c_{i.})} \tag{8}$$

Table 5. Day 1 results where the max values for each metric is highlighted in bold.

	ACC	MAP	MAR	CBA	IAM
Classifier-Sampling	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
RF- bal	0.39 (0.02)	0.34 (0.02)	**0.33 (0.01)**	0.30 (0.02)	−0.41 (0.02)
RF – ROS	0.37 (0.04)	0.32 (0.03)	**0.33 (0.01)**	**0.31 (0.01)**	−0.38 (0.01)
RF – SM	0.36 (0.04)	0.31 (0.03)	**0.33 (0.01)**	0.30 (0.02)	−0.41 (0.03)
GBM – No	**0.40 (0.01)**	**0.37 (0.01)**	0.32 (0.01)	0.28 (0.02)	−0.44 (0.02)
GBM – ROS	**0.40 (0.02)**	0.35 (0.02)	0.32 (0.02)	0.30 (0.02)	−0.40 (0.01)
GBM – SM	0.38 (0.04)	0.33 (0.02)	**0.33 (0.02)**	**0.31 (0.02)**	**−0.37 (0.01)**
NN – No	0.39 (0.02)	0.35 (0.02)	0.28 (0.03)	0.20 (0.03)	−0.60 (0.01)
NN – ROS	0.31 (0.04)	0.29 (0.03)	**0.33 (0.01)**	0.22 (0.02)	−0.57 (0.01)
NN – SM	0.33 (0.04)	0.29 (0.03)	0.31 (0.02)	0.28 (0.01)	−0.49 (0.01)

Table 6. Day 2 results where the max values for each metric is highlighted in bold.

	ACC	MAP	MAR	CBA	IAM
Classifier-Sampling	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
RF- bal	0.54 (0.01)	**0.47 (0.02)**	0.46 (0.02)	0.44 (0.02)	−0.13 (0.02)
RF – ROS	0.53 (0.02)	0.45 (0.02)	0.46 (0.02)	0.43 (0.02)	−0.14 (0.02)
RF – SM	0.53 (0.02)	0.46 (0.02)	**0.48 (0.01)**	0.44 (0.01)	**−0.11 (0.01)**
GBM – No	0.54 (0.01)	**0.47 (0.01)**	0.44 (0.02)	0.42 (0.02)	−0.16 (0.01)
GBM – ROS	0.53 (0.01)	0.46 (0.01)	0.45 (0.02)	0.43 (0.01)	−0.14 (0.01)
GBM – SM	0.53 (0.01)	0.46 (0.02)	0.47 (0.01)	**0.45 (0.02)**	**−0.11 (0.01)**
NN – No	**0.56 (0.02)**	**0.47 (0.01)**	0.45 (0.02)	0.41 (0.02)	−0.19 (0.01)
NN – ROS	0.51 (0.02)	0.43 (0.02)	**0.48 (0.01)**	0.39 (0.03)	−0.22 (0.02)
NN – SM	0.53 (0.01)	0.46 (0.01)	**0.48 (0.01)**	0.43 (0.01)	**−0.11 (0.01)**

Table 7. Day 3 results where the max values for each metric is highlighted in bold.

	ACC	MAP	MAR	CBA	IAM
Classifier-Sampling	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
RF- bal	**0.60 (0.01)**	**0.50 (0.02)**	0.49 (0.02)	0.46 (0.01)	−0.09 (0.02)
RF – ROS	0.58 (0.02)	0.48 (0.02)	0.49 (0.01)	0.45 (0.02)	−0.09 (0.02)
RF – SM	0.57 (0.02)	0.49 (0.02)	**0.50 (0.01)**	0.46 (0.01)	−0.08 (0.02)
GBM – No	0.59 (0.01)	0.49 (0.02)	0.47 (0.02)	0.44 (0.02)	−0.12 (0.02)
GBM – ROS	0.59 (0.01)	0.49 (0.02)	0.47 (0.01)	0.44 (0.02)	−0.12 (0.02)
GBM – SM	0.58 (0.01)	**0.50 (0.02)**	**0.50 (0.01)**	**0.47 (0.01)**	**−0.04 (0.02)**
NN – No	0.59 (0.01)	0.45 (0.02)	0.46 (0.0)	0.41 (0.02)	−0.18 (0.02)
NN – ROS	0.54 (0.01)	0.44 (0.02)	0.47 (0.01)	0.41 (0.02)	−0.18 (0.02)
NN – SM	0.56 (0.02)	0.48 (0.02)	**0.50 (0.01)**	0.44 (0.02)	−0.11 (0.02)

Table 8. Day 4 results where the max values for each metric is highlighted in bold.

	ACC	MAP	MAR	CBA	IAM
Classifier-Sampling	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
RF- bal	**0.57 (0.01)**	0.48 (0.02)	0.48 (0.02)	0.43 (0.02)	−0.13 (0.02)
RF – ROS	0.54 (0.02)	0.47 (0.02)	0.47 (0.02)	0.43 (0.02)	−0.14 (0.02)
RF – SM	0.55 (0.02)	**0.50 (0.01)**	**0.49 (0.01)**	**0.45 (0.01)**	−0.10 (0.01)
GBM – No	0.56 (0.01)	0.48 (0.01)	0.46 (0.01)	0.43 (0.01)	−0.14 (0.01)
GBM – ROS	0.56 (0.01)	0.48 (0.02)	0.46 (0.01)	0.43 (0.01)	−0.14 (0.01)
GBM – SM	0.56 (0.01)	**0.50 (0.01)**	**0.49 (0.01)**	**0.45 (0.01)**	**−0.05 (0.01)**
NN – No	0.55 (0.01)	0.46 (0.01)	0.45 (0.01)	0.41 (0.02)	−0.19 (0.01)
NN – ROS	0.50 (0.02)	0.42 (0.02)	0.43 (0.02)	0.38 (0.01)	−0.24 (0.02)
NN – SM	0.51 (0.02)	0.43 (0.02)	0.44 (0.02)	0.39 (0.01)	−0.22 (0.02)

Table 9. The CBA results with different number of features in Phase III.

a. Top 5	day 1 (reduction)	day 2 (reduction)	day 3 (reduction)	day 4 (reduction)
RF	0.29 (−0.06)	0.40 (−0.09)	0.43 (−0.07)	0.43 (−0.02)
GBM	**0.29 (−0.06)**	**0.41 (−0.09)**	**0.43 (−0.09)**	**0.43 (−0.04)**
NN-GA	0.24 (−0.14)	0.40 (−0.05)	0.40 (−0.07)	0.35 (−0.08)
b. Top 10				
RF	0.30 (−0.03)	0.42 (−0.05)	0.45 (−0.02)	0.43 (−0.04)
GBM	**0.30 (−0.03)**	**0.42 (−0.06)**	**0.45 (−0.04)**	**0.44 (−0.02)**
NN-GA	0.25 (−0.1)	0.41 (−0.05)	0.42 (−0.05)	0.37 (−0.05)
c. Top 15				
RF	0.31	0.44	0.46	0.44
GBM	**0.31**	**0.45**	**0.47**	**0.45**
NN-GA	0.28	0.42 (−0.02)	0.43 (−0.02)	0.38 (−0.02)

4. Results and discussions

Table 5 to 8 have summarised the results for Phase II. The mean and standard deviation values for all five metrics of accuracy are reported. Each algorithm is experimented under three different modes: with no resampling, with ROS, and with SMOTE. For no-resampling mode, we take advantage of balancing the weights of imbalanced classes within the loss function. Those experiments are presented with "-bal" suffix. Under the balanced mode, the weights are adjusted inversely proportional to class frequencies in the input data (King & Zeng, 2001). The KNN, RR, LR, SVM and RF algorithms are trained with balancing the weights in loss function. Table 5 to 8 have

Table 10. Feature importance for the GBM classifier with CBA.

	day 1	day 2	day 3	day 4
1	age	MME D1	MME D2	MME D3
2	BMI	los	MME D1	MME D2
3	income err	age	los	los
4	income	BMI	age	MME D1
5	los	income err	BMI	age
6	surgeon	income	income	BMI
7	zip code	surgeon	income err	income err
8	insurance	insurance	surgeon	income
9	procedure	smoking	insurance	insurance
10	smoking	procedure	smoking	surgeon
11	race	asa	procedure	asa
12	asa	zip code	asa	procedure
13	religion	e-proc	zip code	smoking
14	sex	race	race	zip code
15	e-proc	religion	dch_disp	race
16	dch_disp	dch_disp	sex	sex
17	surg_hist	sex	e-proc	dch_disp
18		surg_hist	religion	e-proc
19			surg_hist	religion
20				surg-hist

summarised the results. The accuracy values of KNN, RR, LR, SVM are below the threshold (25%) in the preliminary analysis (0.22, 0.19, 0.23, 0.22, respectively, for the CBA), hence they are discarded from the feature selection and their results are not shown.

As shown, we have reported five accuracy metrics for group comparison and highlighted the best values in bold, but we primarily use the CBA metric for model selection as it is always lower than MAP, MAR and F-score and is consistent with ACC.

4.1. Discussions

Based on the obtained results from Table 5 to 8, we have listed the important insights derived from the results associated to each day below.

Insights for day 1:

- The GBM in the "SMOTE" mode has outperformed all the other methods with mean accuracy values of 0.31 for the CBA metric. Note that the GBA in the "ROS" mode may offer higher ACC value, but due to the imbalance state of the dataset, it is deemed misleading. That conclusion is confirmed by the IAM as the bottom-line metric.

Insights for day 2:

(1) The GBM in the "SMOTE" mode has outperformed all the other methods with mean accuracy values of 0.45 for the CBA metric which is confirmed by the IAM.

Insights for day 3:

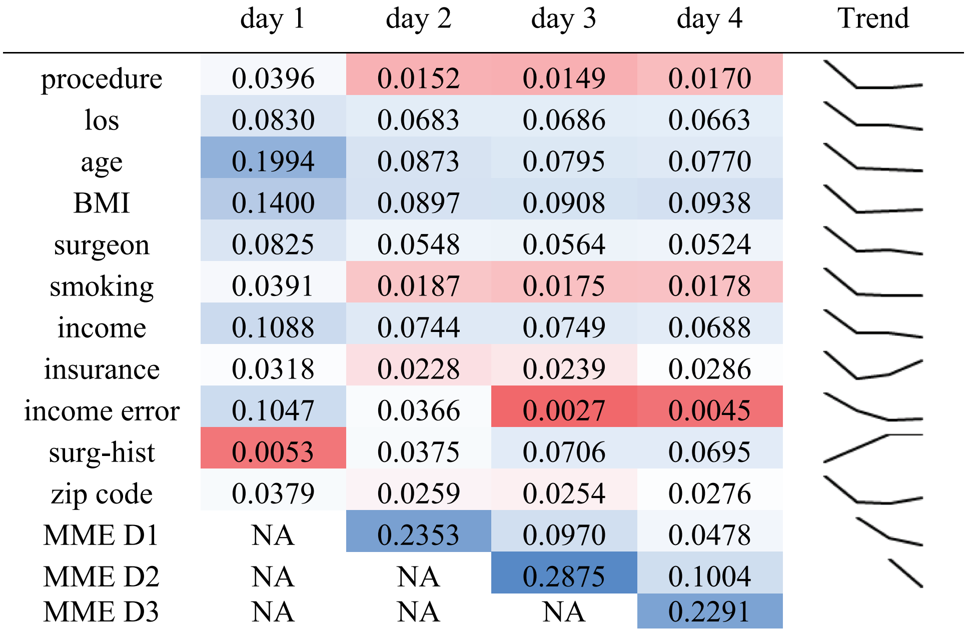
(1) The GBM in the "SMOTE" mode has outperformed all the other methods with mean accuracy value of 0.47 for the CBA metric which is confirmed by the IAM.

Insights for day 4:

(1) The GBM in the "SMOTE" mode has outperformed all the other methods with mean accuracy value of 0.45 for the CBA metric which is confirmed by the IAM.
(2) The "SMOTE" mode has dominated all other modes in the GBM.

Conclusion from all days to assist the surgeons

(1) The GBM in the "SMOTE" mode has outperformed all the classifiers in all 4 days based on the CBA measure. The conclusions are confirmed by the IAM as well.
(2) The results for day 3 have the highest accuracy values and day 1 has the lowest. For day 3, we hypothesise that the promising accuracy values



	day 1	day 2	day 3	day 4	Trend
procedure	0.0396	0.0152	0.0149	0.0170	
los	0.0830	0.0683	0.0686	0.0663	
age	0.1994	0.0873	0.0795	0.0770	
BMI	0.1400	0.0897	0.0908	0.0938	
surgeon	0.0825	0.0548	0.0564	0.0524	
smoking	0.0391	0.0187	0.0175	0.0178	
income	0.1088	0.0744	0.0749	0.0688	
insurance	0.0318	0.0228	0.0239	0.0286	
income error	0.1047	0.0366	0.0027	0.0045	
surg-hist	0.0053	0.0375	0.0706	0.0695	
zip code	0.0379	0.0259	0.0254	0.0276	
MME D1	NA	0.2353	0.0970	0.0478	
MME D2	NA	NA	0.2875	0.1004	
MME D3	NA	NA	NA	0.2291	

Figure 2. Top 10 feature importance obtained from the GBM classifier.

Table 11. The LR model coefficients for range (>120) against the rest of the ranges.

Rank	day 1		day 2		day 3		day 4	
1	age	−6.8034	MME D1	1.9764	income	1.6906	MME D3	1.6778
2	BMI	1.7780	age	1.1610	MME D2	1.4331	surg-hist	1.6678
3	income	1.0087	income	0.8746	asa	1.2260	income err	1.4886
4	Income err	0.6778	income err	0.7113	MME D1	1.2017	BMI	0.9240
5	los	0.3089	surg-hist	0.6476	age	1.0006	income	0.7866
6	smoking	0.2018	sex	0.4081	surg-hist	0.6699	MME D2	0.6986
7	procedure	0.2009	los	0.3389	sex	0.6853	age	0.6899
8	sex	0.1689	asa	0.2613	BMI	0.5132	asa	0.6511
9	surg-hist	0.1283	bmi	0.2034	procedure	0.3781	sex	0.4949
10	asa	0.0869	smoking	0.0963	los	0.1820	los	0.3751
11	religion	0.0633	race	0.0843	race	0.1799	insurance	0.2560
12	dis-disp	0.0629	zipcode	0.0281	smoking	0.1581	MME D1	0.2423
13	insurance	0.0361	dch-disp	0.0212	e-proc	0.1024	procedure	0.2064
14	e-proc	0.0301	surgeon	0.0197	income err	0.0617	religion	0.2028
15	race	0.0254	insurance	0.0185	religion	0.0486	race	0.1688
16	zipcode	0.0164	procedure	0.0173	insurance	0.0264	e-proc	0.1048
17	surgeon	0.0011	religion	0.0154	surgeon	0.0191	smoking	0.0481
18			e-proc	0.0100	zipcode	0.0171	zipcode	0.0363
19					dch-disp	0.0097	surgeon	0.0242
20							dch-disp	0.0203

are due to the availability of day 1 and day 2 prescribed dosage data. With that logic, day 4 results should outdo those of day 3 but the opposite is indeed true. We surmise that it is caused by the smaller sample size for day 4 (less than half of day 3 sample).

It is also interesting to note that the mean value of IAM is always negative. That is, the number of times the classifiers classified the instances correctly on average is lower than the number of the times they classified them incorrectly for each class (bottom-line metric).

With Phase II concluded, the RF, GBM and NN advance (acceptable accuracy) to Phase III, the feature selection. In tree-based ensemble methods, RF and GBM, the relative depth of a feature used as a decision node in a tree is usually considered to assess the relative importance of that feature with respect to the predictability of the target variable. We follow the extended measure used in (Pedregosa et al., 2011) based on (Friedman, 2001; Louppe, 2014) for RF and GBM where the fraction of samples a feature contributes to is combined with the decrease in impurity (Gini importance) to define feature importance. The pseudo code is given in Table A1 of appendix.

We collect 100 sorted arrays of the RF and GBM algorithms that had given the highest accuracy values. For the NN, we conduct the GA algorithm. Table A2 of Appendix has provided the parameters used in GA-NN method. We run the experiments with top 5, top 10 and top 15 features of each algorithm and summarise the CBA in Table 9. The differences in the accuracy results from Table 5 to 8 (outcomes of the models with all features) are also included. Top 15 results are almost the same as those obtained from all-featured cases implying those discarded could be disregarded from the prescription application. Those features are *discharge disposition* and *surgical history* for day 1, *discharge disposition, patients' sex* and surgical history for day 2, *discharge disposition, sex* and *surgical history for day 2, extended procedure, sex, religion and surgical history* for day 3, and *extended procedure, sex, religion, discharge disposition* and *surgical history* for day 4. Using top 10 features shows some reduction on the accuracy values but the differences are not significant. Using top 5 features also shows some reduction but again the reductions are not significant. It is worth noting that the top 5 features case includes only numerical features which shows the how important those features are in training the classifiers.

Table 10 has shown the sorted features based on their importance in the GBM algorithm and the top 10 are highlighted. As the GBM outperformed the RF and the NN, we have presented the important features in detail only for the GBM. We have provided the sorted features for the RF and the NN-GA in Tables A3 and A4 of Appendix. Figure 2 provides a heatmap that depicts the top 10 important values with different colours for the GBM (from blue (most important) to red (least important)). Based on Figure 2, *age* is the most important feature in day 1 and the prescribed doses from the preceding days are the most important feature for the following days (for instance, MME D1 for day 2, MME D2 for day 3 and so on.). The *age, BMI, income, surgeon, los* and *insurance* variables are very important during day 1 but their importance diminishes in the following days. The highest decline is associated to *age* and the lowest decline is for *los*. That is, the *los* maintains its importance throughout the 4-day regimen. On the other hand, the prescribed doses from the preceding days appear to become very important as they are being considered in the prediction mix. The last column in Figure 2 shows the trend (increasing or decreasing) from day 1 to day 4. Some of the features have high impacts in day 1 but their importance fades away in subsequent days such as *age,* *BMI, procedure, income* and *income error*. Some of the features are less important in day 1 and they become important in subsequent days. Those include *surgical history* and the preceding days prescriptions as being added gradually from day 1 to day 3.

5. Selected features evaluation

The LR model is a statistical model whose outcomes unlike most machine learning algorithms are interpretable. With a sacrifice on model accuracy, the LR models can be used for explanation purposes. We have provided the values of the final coefficients for range >120-against-all ranges in the LR model in Table 11. As all our variables are normalised, we can use the magnitude to explain the importance in the LR model. The positive sign indicates a direct correlation between the feature and the target variable and the negative values indicate the opposite. In day 1, holding all variables except *age* at a fixed value, the odds of getting into range 5 (>120) is exp(−6.80) = 0.001. That is, as the patients get older, the odds of prescribing >120 decreases. Similarly, that number for BMI is 5.87 and for income is 2.77 in day 1. The changes in the odds for the other days can be interpreted in a similar fashion. For all four days, regarding procedure, the odds of prescribing >120 is the highest for total knee, and then shoulder and hip arthroplasty. The odds for the unilateral arthroplasty and revisions procedures are lower. The longer the patients stays, the odds of over utilisation increases (within four-day stay). For the smoking, the odds of over utilisation is the highest for current every day smoker and current some day smoker, and the lowest for former smoker and never smokers. The odds of over utilisation is higher in the male patients than the females patients, and in the patients who have had surgeries in the past (surg-hist > 0). The interpretations in greater details for day 1 have been presented in Table 12. For the following days, the interpretations would be similar.

Table 12. Interpretation of the top 10 LR coefficients for day 1.

Rank	Feature	Coefficient	Odds of prescribing >120 MME
1	age	−6.8034	Increases as the age decreases
2	BMI	1.7780	Increases as the BMI increases
3	income	1.0087	Increases as the income increases
4	Income err	0.6778	Increases as the income err increases
5	los	0.3089	Increases as the length of stay increases
6	smoking	0.2018	Increases from non-smokers to every day smokers
7	procedure	0.2009	Increases from revisions and unilaterals to total hip, shoulder and knee
8	sex	0.1689	Increases from women to men
9	surg-hist	0.1283	Increases from a single surgery to multiple surgery (1–3)
10	asa	0.0869	Increases from 1 to 4

6. Conclusions, limitations and future work

The main objective of this research was to offer a ML-based model to identify the varying effects of factors that are related with short-term opioid prescription amount, after orthopaedic surgery. A multi-class prediction problem was derived from the dataset and multiple classifiers were trained to perform the classification task.

Specific conclusions for the machine learning aspects of the model can be summarised as follows:

(1) The GBM in the “SMOTE” mode has outperformed all the algorithms at all four days based on the CBA metric.
(2) The feature importance results for all three methods that survived phase II showed five to seven common features among their top 10 most important features. Patient’s age, BMI, length of hospital stay, surgical procedure, income and prior prescriptions are among the most important features impacting the opioid prescription in orthopaedics patients.
(3) The feature selection can remove the noisy variables and improve the quality of fit. We showed that training the models with top 10 most important features makes the models simpler which leads to faster execution. The accuracy reduction was also not significant.

Specific conclusions for the clinicians can be summarised as follows:

(1) The trained GBM model can readily be used to classify the postoperative orthopaedics patients into one of the five prescription ranges.
(2) For day 1, the age, and for the following days, the doses prescribed in the previous day are the most important features.
(3) Based on the logistic regression analysis, in day 1, as the patients get older, the odds of prescribing >120 MME range decreases. For higher BMI and income values, the odds increase. For procedures, the odds are the highest for total knee, and then shoulder and hip arthroplasty. The odds for the unilateral arthroplasty and revisions procedures are lower. The revision procedures depict lower risk. The longer the patients stays, the odds of prescribing higher opioid increases. For patients who smoke, the odds of over utilisation is the highest for current every day smoker and current some day smoker, and the lowest for former smoker and never smokers. The odds of higher doses are higher in the male patients than the females patients, and in the patients who have had multiple surgeries in the past (surg-hist > 0).

Additionally, we should note that one limitation in our study is the absence of the information on the opioid dependency state of the patients (opioid naive vs opioid dependent) which would be impactful (Mitra & Sinatra, 2004). Considering that variable in future research could improve the attractiveness of our proposition for adoption.

In addition, the measures of accuracy for day 1 are not particularly high. We believe that the accuracy values could be enhanced if a larger dataset possibly from other institutions and the data on IV intakes in *day 0* were available. This along with the information on opioid dependency could enhance our work.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Science+ Business Media.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cook, C. E., Rhon, D. I., Lewis, B. D., & George, S. Z. (2017). Post-operative opioid pain management patterns for patients who receive hip surgery. *Substance Abuse Treatment, Prevention, and Policy*, *12*(1), 14. https://doi.org/10.1186/s13011-017-0094-5

Dag, A., Oztekin, A., Yucel, A., Bulur, S., & Megahed, F. M. (2017). Predicting heart transplantation outcomes through data analytics. *Decision Support Systems*, *94*, 42–52. https://doi.org/10.1016/j.dss.2016.10.005

Dag, A., Topuz, K., Oztekin, A., Bulur, S., & Megahed, F. M. (2016). A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decision Support Systems*, *86*, 1–12. https://doi.org/10.1016/j.dss.2016.02.007

Dolatsara, H. A., Chen, Y. J., Evans, C., Gupta, A., & Megahed, F. M. (2020). A two-stage machine learning framework to predict heart transplantation survival probabilities over time with a monotonic probability constraint. *Decision Support Systems*, *137*, 113363. https://doi.org/10.1016/j.dss.2020.113363

Dowell, D., Haegerich, T. M., & Chou, R. (2016). CDC guideline for prescribing opioids for chronic pain—United States, 2016. *JAMA*, *315*(15), 1624–1645. https://doi.org/10.1001/jama.2016.1464

Dursun, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5).

Goesling, J., Moser, S. E., Zaidi, B., Hassett, A. L., Hilliard, P., Hallstrom, B., et al. (2016). Trends and predictors of opioid use following total knee and total hip arthroplasty. *Pain*, *157*(6), 1259. https://doi.org/10.1097/j.pain.0000000000000516

Kee, J. R., Smith, R. G., & Barnes, C. L. (2016). Recognizing and reducing the risk of opioid misuse in orthopedic practice. *Journal of Surgical Orthopedic Advances*, *25*(4), 238–243.

Kim, S. C., Choudhry, N., Franklin, J. M., Bykov, K., Eikermann, M., Lii, J., et al. (2017). Patterns and predictors of persistent opioid use following hip or knee arthroplasty. *Osteoarthritis and Cartilage*, *25*(9), 1399–1406. https://doi.org/10.1016/j.joca.2017.04.002

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*(2), 1137–1145.

Kroenke, K., Alford, D. P., Argoff, C., Canlas, B., Covington, E., Frank, J. W., et al. (2019). Challenges with implementing the centers for disease control and prevention opioid guideline: A consensus panel report. *Pain Medicine*, *20*(4), 724–735. https://doi.org/10.1093/pm/pny307

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kusiak, A., Bradley, D., & Shah, S. (2005). Predicting survival time for kidney dialysis patients: A data mining approach. *Computers in Biology and Medicine*, *35*(4), 311–327. https://doi.org/10.1016/j.compbiomed.2004.02.004

Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, *6*(5), 267–281. https://doi.org/10.1002/cem.1180060506

Lo-Ciganic, W. H., Huang, J. L., Zhang, H. H., Weiss, J. C., Wu, Y., Kwoh, C. K., et al. (2019). Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA Network Open*, *2*(3), e190968–e190968. https://doi.org/10.1001/jamanetworkopen.2019.0968

Louppe, G. (2014). *Understanding random forests: from theory to practice* [PhD Thesis]. U. of Liege.

Mitra, S., & Sinatra, R. S. (2004). Perioperative management of acute pain in the opioid-dependent patient. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, *101*(1), 212–227. https://doi.org/10.1097/00000542-200407000-00032

Morris, B. J., Mir, H. R., & Mir. (2015). The opioid epidemic: Impact on orthopedic surgery. *JAAOS-Journal of the American Academy of Orthopedic Surgeons*, *23*(5), 267–271. https://doi.org/10.5435/JAAOS-D-14-00163

Mortaz, E. (2020). Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems*, *210*, 106490. https://doi.org/10.1016/j.knosys.2020.106490

Mosley, L. (2013). *A balanced approach to the multi-class imbalance problem*. http://lib.dr.iastate.edu

Nasir, M., South-Winter, C., Ragothaman, S., & Dag, A. (2019). *A comparative data analytic approach to construct a risk trade-off for cardiac patients' re-admissions*. Industrial Management & Data Systems.

Nasir, M., Summerfield, N., Dag, A., & Oztekin, A. (2020). A service analytic approach to studying patient no-shows. *Service Business*, *14*(2), 287–313. https://doi.org/10.1007/s11628-020-00415-8

National Academies of Sciences, Engineering, and Medicine. (2017). *Pain management and the opioid epidemic: Balancing societal and individual benefits and risks of prescription opioid use*. National Academies Press.

Oztekin, A., Al-Ebbini, L., Sevkli, Z., & Delen, D. (2018). A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. *European Journal of Operational Research*, *266*(2), 639–651. https://doi.org/10.1016/j.ejor.2017.09.034

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Prieto, J. T., Scott, K., McEwen, D., Podewils, L. J., Al-Tayyib, A., Robinson, J., et al. (2020). The detection of opioid misuse and heroin use from paramedic response documentation: Machine learning for improved surveillance. *Journal of Medical Internet Research*, *22*(1), e15645. https://doi.org/10.2196/15645

Rhon, D. I. (2018). Predictors of chronic prescription opioid use after orthopedic surgery: Derivation of a clinical prediction rule. *Perioperative Medicine*, *7*(1), 25. https://doi.org/10.1186/s13741-018-0105-8

Rudd, R. A. (2016). Increases in drug and opioid-involved overdose deaths—United States, 2010–2015. *MMWR. Morbidity and Mortality Weekly Report*, *65*(5051), 1445–1452. https://doi.org/10.15585/mmwr.mm655051e1

Schoenfeld, A. J., Belmont, J. P. J., Blucher, J. A., Jiang, W., Chaudhary, M. A., Koehlmoos, T., et al. (2018). Sustained preoperative opioid use is a predictor of continued use following spine surgery. *JBJS*, *100*(11), 914–921. https://doi.org/10.2106/JBJS.17.00862

Shreve, J., Schneider, H., & Soysal, O. (2011). A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. *Decision Support Systems*, *52*(1), 247–257. https://doi.org/10.1016/j.dss.2011.08.001

Simsek, S., Kursuncu, U., Kibis, E., AnisAbdellatif, M., & Dag, A. (2020). A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. *Expert Systems with Applications*, *139*, 112863. https://doi.org/10.1016/j.eswa.2019.112863

Simsek, S., Tiahrt, T., & Dag, A. (2020). Stratifying no-show patients into multiple risk groups via a holistic data analytics-based framework. *Decision Support Systems*, *132*, 113269. https://doi.org/10.1016/j.dss.2020.113269

Sing, D. C., Barry, J. J., Cheah, J. W., Vail, T. P., & Hansen, E. N. (2016). Long-acting opioid use independently predicts perioperative complication in total joint arthroplasty. *The Journal of Arthroplasty*, *31*(9), 170–174. https://doi.org/10.1016/j.arth.2016.02.068

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems*, *106*, 97–109. https://doi.org/10.1016/j.dss.2017.12.004

Washington State Agency Medical Directors' Group. *AMDG 2015 interagency guideline on prescribing opioids for pain*. http://www.agencymeddirectors.wa.gov/guidelines.asp

Zarling, B. J., Yokhana, S. S., Herzog, D. T., & Markel, D. C. (2016). Preoperative and postoperative opiate use by the arthroplasty patient. *The Journal of Arthroplasty*, *31*(10), 2081–2084. https://doi.org/10.1016/j.arth.2016.03.061

Zheng, B., Zhang, J., Yoon, S. W., Lam, S. S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, *42*(20), 7110–7120. https://doi.org/10.1016/j.eswa.2015.04.066

Zywiel, M. G., Stroh, D. A., Lee, S. Y., Bonutti, P. M., & Mont, M. A. (2011). Chronic opioid use prior to total knee arthroplasty. *JBJS*, *93*(21), 1988–1993. https://doi.org/10.2106/JBJS.J.01473

Appendix

Table A1. Feature importance code for ensemble methods.

```
Total sum = empty array (size = number of trees)
   for tree in trees:
   stage sum = sum (importance array at each tree for each stage)/size(stage)
   Total sum + = stage sum
   " The stage represents the one-vs-all scheme in RF/GBM)"
   Importance array = Total sum/size(trees)
   return importance array
   Importance array at each tree:
   while node ! = leaf node:
   importance data[node. feature] + = (
   node. n-samples × node. impurity -
   left node. n-samples × left. impurity -
   right node. n-samples × right. impurity
   node + = 1
   importance/ = number of observations
   return importance
   impurity: Gini index.
```

Table A2. The GA parameters.

population size	max-gen	selection	crossover method (rate)	mutation rate	children	α
50	100	Tournament	Single-point (0.1)	0.05	4	0.9

Table A3. RF feature selection results.

Rank	day 1	day 2	day 3	day 4
1	BMI	MME D1	MME D2	MME D3
2	age	BMI	BMI	BMI
3	income err	age	los	age
4	income	income err	age	income error
5	surgeon	income	income	income
6	los	surgeon	income err	MME D2
7	zip code	los	surgeon	los
8	procedure	zip code	MME D1	surgeon
9	insurance	insurance	zip code	MME D1
10	smoking	procedure	insurance	procedure
11	religion	smoking	procedure	zip code
12	race	religion	smoking	dch_disp
13	asa	e-proc	religion	smoking
14	dch_disp	dch_disp	dch_disp	race
15	e-proc	asa	asa	religion
16	sex	race	race	asa
17	surg_hist	sex	e-proc	sex
18		surg_hist	sex	insurance
19			surg_hist	surg_hist
20				e-proc

Table A4. NN-GA feature selection results.

Rank	day 1	day 2	day 3	day 4
1	BMI	MME D1	MME D2	MME D3
2	age	los	los	surgeon
3	los	age	BMI	los
4	procedure	BMI	age	BMI
5	income	dch_disp	asa	age
6	sex	income	dch_disp	asa
7	asa	asa	income	insurance
8	religion	insurance	MME D1	income
9	surgeon	smoking	insurance	MME D2
10	surg_hist	surg_hist	smoking	dch_disp
11	Income err	e-procedure	surg_hist	smoking
12	zip code	sex	e-procedure	surg_hist
13	race	religion	religion	e-procedure
14	smoking	Income err	sex	zip code
15	e-procedure	zip code	Income err	sex
16	dch_disp	procedure	zip code	Income err
17	insurance	race	procedure	religion
18		surgeon	race	procedure
19			surgeon	race
20				MME D1