Pain Medicine

Pain Medicine 2015; 16: 1386–1401
Wiley Periodicals, Inc.

ACUTE & PERIOPERATIVE PAIN SECTION

Original Research Article

Teaching a Machine to Feel Postoperative Pain: Combining High-Dimensional Clinical Data with Machine Learning Algorithms to Forecast Acute Postoperative Pain

Patrick J. Tighe, MD, MS,*
Christopher A. Harle, PhD,†
Robert W. Hurley, MD, PhD,* Haldun Aytug, PhD,‡
Andre P. Boezaart, MD, PhD,*,§
and Roger B. Fillingim, PhD¶

*Department of Anesthesiology, University of Florida College of Medicine, Gainesville, Florida, USA; †Department of Health Services Research, Management and Policy, College of Public Health and Health Professions, Gainesville, Florida, USA; ‡Department of Information Systems and Operations Management, Warrington College of Business Administration, Gainesville, Florida, USA; §Department of Orthopaedic Surgery and Rehabilitation, University of Florida College of Medicine, Gainesville, Florida, USA; ¶Department of Community Dentistry and Behavioral Science, University of Florida College of Dentistry, Gainesville, Florida, USA

Reprint requests to: Patrick Tighe, MD, MS, Department of Anesthesiology, University of Florida, College of Medicine, PO Box 100254, 1600 SW Archer Road, Gainesville, FL 32610, USA. Tel: 352-273-7844; Fax: 352-392-7029; E-mail: ptighe@anest.ufl.edu.

Disclosure: Department/institution to which this work is attributed: Departments of Anesthesiology; Information Systems and Operations Management, Warrington College of Business Administration; and Community Dentistry and Behavioral Science, University of Florida, Gainesville, Florida, USA.

Funding sources: Patrick J. Tighe is funded by an NIH grant (no. K23GM102697) and support for Christopher A. Harle on this study was provided in part by grants from the NIH (NCATS) UL1TR000064 and CTSA KL2TR000065.

Conflict of interest: The authors have no conflicts of interests to report.

Abstract

Background. Given their ability to process highly dimensional datasets with hundreds of variables, machine learning algorithms may offer one solution to the vexing challenge of predicting postoperative pain.

Methods. Here, we report on the application of machine learning algorithms to predict postoperative pain outcomes in a retrospective cohort of 8,071 surgical patients using 796 clinical variables. Five algorithms were compared in terms of their ability to forecast moderate to severe postoperative pain: Least Absolute Shrinkage and Selection Operator (LASSO), gradient-boosted decision tree, support vector machine, neural network, and k-nearest neighbor (k-NN), with logistic regression included for baseline comparison.

Results. In forecasting moderate to severe postoperative pain for postoperative day (POD) 1, the LASSO algorithm, using all 796 variables, had the highest accuracy with an area under the receiver-operating curve (ROC) of 0.704. Next, the gradient-boosted decision tree had an ROC of 0.665 and the k-NN algorithm had an ROC of 0.643. For POD 3, the LASSO algorithm, using all variables, again had the highest accuracy, with an ROC of 0.727. Logistic regression had a lower ROC of 0.5 for predicting pain outcomes on POD 1 and 3.

Conclusions. Machine learning algorithms, when combined with complex and heterogeneous data

from electronic medical record systems, can forecast acute postoperative pain outcomes with accuracies similar to methods that rely only on variables specifically collected for pain outcome prediction.

Key Words. Machine Learning; Algorithm; Postoperative Pain; Pain Prediction

Introduction

Over 60% of surgical patients suffer from moderate to severe acute postoperative pain, and this pain has been associated with the development of chronic postsurgical pain [1,2]. Mounting evidence points to the importance of establishing preemptive, and even preventative, analgesia whenever possible before the onset of surgical stimulus [3,4]. However, many preemptive and preventative analgesic interventions can carry considerable side effects, such as bleeding or major adverse cardiac events with nonsteroidal antiinflammatories and sedation with gabapentinoids. Therefore, the ability to predict which patients are more likely to suffer from moderate to severe acute postoperative pain would permit targeting of perioperative analgesic therapies in a manner that optimizes the risk to benefit ratio.

Accurate postoperative pain prediction has been the topic of research for over a century [5]. Although previous efforts that used logistic regression have highlighted potential risk factors for severe postoperative pain, these approaches are limited [6,7]. For instance, logistic regression approaches are unable to incorporate the rapidly expanding set of available clinical data, let alone the genetic, proteomic, and metabolomic data expected to be available for clinical decision support systems in the near future [6–13]. Pragmatically, such approaches also require regular updating to remain relevant to modern practice. Thus, new methods are needed that incorporate the potential predictive power of the myriad data elements being routinely collected. Moreover, new methods are needed that can automatically select the most useful variables and develop and validate prediction algorithms to stay current with current clinical practice.

Machine learning classifiers are algorithms that can autonomously integrate and learn from complex datasets with many hundreds of variables. Therefore, machine learning classifiers may offer a solution to the vexing challenge of predicting postoperative pain [14]. These algorithms use a variety of mathematical approaches and are often more computationally efficient and accurate when using very large datasets with complex distributions that do not conform to the assumptions of parametric methods like logistic regression [15–18]. Machine learning classifiers have already been successfully applied to many prediction problems, including crime prevention, handwriting recognition, fraud detection, and email spam filtering [19–22]. Furthermore, the recent focus on the adoption and meaningful use of electronic medical records (EMR) has led to massive clinical datasets comprising variables collected by healthcare providers during the course of a patient's hospitalization [23–25]. Machine learning approaches have the potential to leverage this clinical "Big Data" to create more accurate and automated predictions of postoperative pain.

Here, we explore the application of machine learning algorithms to analyzing the highly complex data available in the preoperative period to accurately predict acute postoperative pain. The primary goal was to test the feasibility of an automated machine learning process to collect, prepare, and classify preoperative patient data from an EMR and determine whether a patient was at risk for moderate to severe postoperative pain. The secondary goal was to determine the proportion of at-risk patients that could be reliably identified with machine learning algorithms. Together, these aims lay a foundation for the future incorporation of highly complex clinical features into a clinical decision support system that predicts which patients are at risk of postoperative pain and guides clinicians toward the safest and most effective preemptive and preventative analgesic interventions.

Materials and Methods

Study Design

This study was approved by the Institutional Review Board (IRB 354-2012) at the University of Florida and was a retrospective cohort study of surgical patients undergoing nonobstetric, nonambulatory surgical procedures over a 1-year time period from May 2011 to May 2012 at a large tertiary-care teaching hospital.

Description of Dataset

Surgical case data were obtained from the University of Florida's Integrated Data Repository, which is a large database of patient demographic characteristics and care data obtained from the university health system's (UF Health) EMR system. Subjects were patients aged 21 and over undergoing nonambulatory surgery at UF Health over a 1-year period beginning May 2011. Surgical case exclusion criteria included obstetric surgery, as well as patients who received multiple separate surgeries within the study period to avoid contamination of pain scores from the effects of surgeries preceding or following the case of interest. Results were reported in accordance with the STROBE criteria for cohort studies. (http://www.strobe-statement.org/fileadmin/Strobe/uploads/checklists/STROBE_checklist_v4_cohort.pdf)

Description of Outcomes

All pain scores were documented by clinical staff using the numeric rating scale (NRS) on an 11-point system ranging from 0 to 10, where zero represents no pain and 10 the worst pain imaginable. Pain scores were

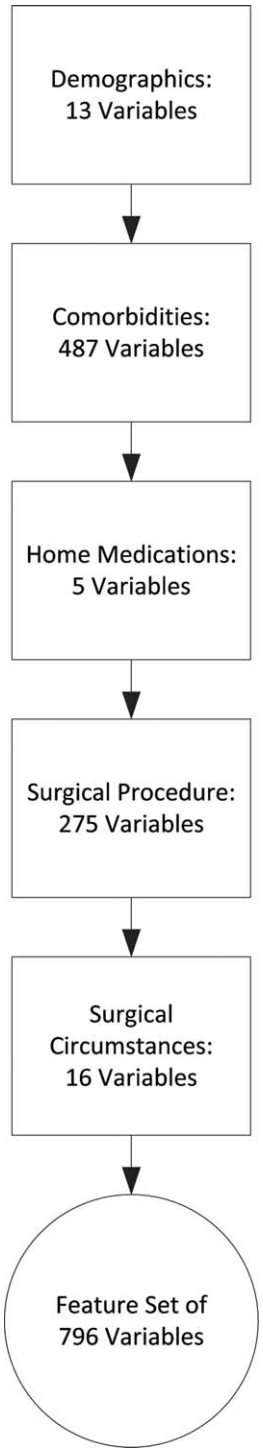

Figure 1 Loading of variables into machine learning classifier pipeline. Variables were included using a staged approach for demographics, comorbidities, home medications, surgical procedure, and the circumstances of surgery.

recorded every 4 hour per nursing protocol, with a repeat query within 1 hour after administration of analgesic medications for breakthrough pain. When the clinical staff documented a pain score as "patient asleep," the pain score was converted to a missing value rather than 0/10 to account for the fact that some patients had received additional sedatives that may have facilitated sleep despite ongoing pain. All pain scores were recorded with a corresponding date/time stamp, as were the start and end times of the related surgical procedure. End of surgery times generally reflected the closure of skin and emergence from anesthesia.

Two outcomes were defined: the presence or absence of a moderate (NRS score of 4–6) to severe (NRS score of 7–10) maximum pain score on postoperative day (POD) 1 and on POD 3. POD 1 and 3 were selected to address challenges with the early adaptive response of the healthcare system to address patient needs on POD1, as well as patients with refractory pain on POD3, despite the theoretical escalation of pain therapies for at least 48 hour after surgery [26].

Description of Variables

Predictions were rendered based on 796 variables (Figure 1). This compares to the use of only 24 or fewer variables in previous work [26,27]. Demographic data included age, gender, body mass index, ethnicity, insurance/payer, and marital status. Binary variables were defined based on the presence or absence of home use of opioids, nonsteroidal antiinflammatory drugs (NSAIDs), muscle relaxants, benzodiazepines, and amine reuptake inhibitors. Medications were extracted using the World Health Organization pharmaceutical ontology (http://www.whocc.no/atc_ddd_index/).

Patient comorbidity data were prepared by first extracting up to 50 comorbid diagnoses per patient. Diagnoses were recorded using the *International Classification of Disease, 9th edition, Clinical Modification* (ICD-9-CM). Each diagnostic code was also associated with a "present on admission" flag, denoting that the diagnosis was explicitly documented as a diagnosis occurring prior to hospital admission. Also, the ICD-9-CM codes were then converted into a Charlson Comorbidity Index [28]. Separate from the Charlson Comorbidity Index, the total number of comorbid conditions was also calculated. Next, comorbid diagnoses were included in the analysis using 30 binary variables. These categorical variables were defined by the presence or absence of 1 of 30 predefined Agency for Healthcare Research and Quality (AHRQ) comorbidity codes (http://www.ncbi.nlm.nih.gov/pubmed/9431328?dopt=Abstract). Additionally, a parallel and corresponding variable was assigned to each comorbid diagnosis. Each ICD-9-CM diagnosis was recoded as Clinical Categorization Software (CCS) for Services and Procedures diagnosis according to the CCS system (http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp). Finally, for each of the 288 separate CCS diagnoses, the presence or

absence of the diagnosis was arrayed as a binary variable, irrespective of order of entry. Ultimately, an array of 48,787 variables pertaining to established comorbidities was loaded into the machine learning process.

The identities of the surgeon, anesthesiologist, nurse, time of surgery (day of week, weekday versus weekend, normal versus off-hours), postoperative admission versus inpatient status, nerve block status, and emergent versus elective status of the procedure were included and organized into 16 separate variables used to describe the circumstances of the surgery. Types of surgery were identified using current procedural terminology (CPT) codes published by the American Medical Association. Up to 10 CPT codes were included for each patient, and a count of the number of concurrent CPT codes was also included as a covariate. Given the large number of CPT codes, surgeries were grouped into 245 separate categories according to the CCS system, as well as a broader grouping using anatomic location of surgery based on the first one to three digits of the CPT code (http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp). The CCS grouping was performed using a ranked parallel listing of CCS procedure groups as well as a wide array of CCS groups represented as binary flags. Ultimately, 275 variables were included to describe and categorize the type of procedures performed.

Machine Learning Process: Data Preparation

Figure 2 outlines the overall experimental design. First, data were imported as two discrete tables, one including all cases with an outcome (i.e., a valid pain score) on POD1, and a subset of this table for patients who also had an outcome on POD3. The next step in data cleansing was imputation of missing data. Because several of the algorithms would not function if missing values were present, we used a protocol for automated entry of missing data. While this approach inevitably leads to information loss, this step improves the clinical feasibility for implementing an automated clinical decision support system with real-world hospital administrative datasets, which frequently contain missing data. Additionally, this step tested the ability of the analysis to function automatically, such as in a setting where manual cleaning and imputation would be infeasible. For nominal variables, missing entries were imputed using the distribution method, whereby replacement values for a given variable were based on the normalized random percentiles of that variable's distribution. For continuous variables, the median value for a given variable was used for imputation.

Next, we used three levels of interventions to address the risk of overfitting, whereby the model is overcustomized to existing data and less useful for predicting future patient outcomes [27–31]. First, data were partitioned into training (40% of observations), validation (30% of observations), and hold-out testing (30% of observations) partitions. Each partition was stratified on the target outcome so that roughly equivalent proportions of moderate to severe pain outcomes were present in each partition. Second, we included an experiment branch that included an automated variable selection algorithm that selected a subset of variables for use by the algorithms. Third, several of the algorithms tested incorporated regularization features and/or additional cross-validation in their modeling process.

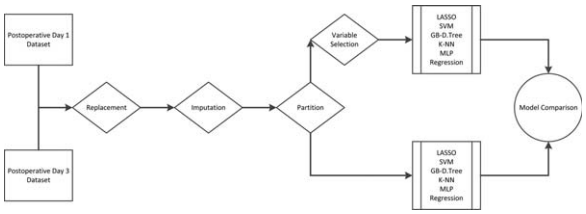

Figure 2 Overview of machine learning classifier pipeline. Separate experiments were conducted for outcomes occurring on POD 1 and 3. Data replacement, imputation, and partitioning were performed using an algorithmic approach. Five machine learning classifiers were tested, along with a standard logistic regression classifier, using the entire set of variables, as well as a reduced set of variables selected via a separate feature set reduction algorithm. LASSO = least absolute shrinkage and selection operator; SVM = support vector machine; GB-D.Tree = gradient-boosting decision tree; k-NN = k-nearest neighbor; MLP = multilayer perceptron.

Description of Algorithms

Five separate algorithms were tested in the classification array: Least Absolute Shrinkage and Selection Operator (LASSO), gradient-boosted decision tree, support vector machine (SVM), neural network, k-nearest neighbor (k-NN), and logistic regression. These algorithms were chosen to represent a wide variety of classification approaches ranging from the classic (logistic regression) to those specifically designed to accommodate highly dimensional data (SVM and LASSO). Details of the selected algorithms, and their implementation, can be found in the Supporting Information technical supplement.

Analysis

Following the training and validation of each algorithm with full and reduced variable sets, algorithm accuracy was compared by examining accuracy in classifying moderate to severe pain in the holdout test data partition [32]. The primary endpoint for comparison of model accuracy was the area under the receiver-operating curve (ROC) [33,34]. Misclassification rates and the

Table 1 Subject demographics

		POD1_NRS_Median_ModSev		POD3_NRS_Median_ModSev	
		No	Yes	No	Yes
Mean age		60.92376	51.59456	61.6973	52.81206
Std Dev age		15.39292	15.69969	15.26899	15.51356
Gender					
Female	*N*	1,821	2,253	1,326	1,253
	% of total	22.56%	27.91%	26.36%	24.91%
Male	*N*	1,983	2,014	1,449	1,003
	% of total	24.57%	24.95%	28.80%	19.94%
Mean BMI		25.71509	25.81831	26.05066	26.32012
Std Dev BMI		11.65954	12.0961	10.5208	11.96438
Charlson Comorbidity Index	Mean	1.202944	0.944692	1.360721	1.070922
	Std Dev	1.21145	1.156419	1.27845	1.177892
Mean number of CPT codes		1.671399	1.738224	1.672072	1.791223
Std Dev number of CPT codes		1.062255	1.178471	1.065549	1.265378
Mean number of comorbidities		9.748686	9.117178	11.00396	10.55408
Std Dev number of comorbidities		6.789532	6.472912	7.058233	6.760371
Timing of surgery					
Elective		3,372	3,742	2,460	1,972
Emergent		34	59	27	45
Urgent		398	466	288	239
Patient admission status					
Inpatient		1,559	1,949	1,205	1,071
Outpatient		202	266	78	104
Present on admission		2,028	2,036	1,483	1,071
Anatomical classification of primary CPT code of surgery					
Auditory		8	8	4	3
Cardiovascular		626	389	528	237
Digestive		812	901	586	472
Endocrine		36	45	21	13
Eye and orbit		3	4	1	3
Female genital		132	162	79	75
Heme and lymphatic		25	27	16	13
Integumentary		185	393	127	216
Male genital		15	17	7	10
Maternity care and delivery		10	13	6	4
Mediastinum and diaphragm		10	6	11	2
Musculoskeletal		798	1,151	618	683

Table 1: *Continued*

	POD1_NRS_Median_ModSev		POD3_NRS_Median_ModSev	
	No	Yes	No	Yes
Nervous	549	593	304	262
Other	141	116	50	40
Pulmonary	171	115	154	63
Reproductive system and intersex	1	1	0	0
Urinary	282	326	263	160
N	3,804	4,267	2,775	2,256

number of wrong classifications were reported to offer clinical context of the ROC [35]. Additionally, we computed error matrices for to determine in which direction the errors were made.

As a secondary endpoint of classifier performance, we reported the cumulative lift for each model [36–38]. Lift measures how many times more likely an algorithm is to include instances of interest (patients with pain in this case) relative to pure chance if we had to choose only a small subsample (i.e., we want the subsample to include as many patients with pain as possible). It is the ratio of the percentage of patients with a high(er) probability of pain as predicted by the model to the percentage of patients with pain in the overall dataset. For example, take the case where an acute pain service could offer only a limited number of nerve blocks each day, such that only 20% of eligible surgical patients could receive a block. Given a predictive model, we assume rightly or wrongly, our best chance of identifying that subpopulation most likely to otherwise suffer from severe pain is to examine the prediction probabilities of the model and pick a sample from the general surgical population that has the top 20% of the predicted probability of suffering from severe acute postoperative pain. A perfect model would fill that entire sample with patients who actually will suffer from severe acute postoperative pain. If the distribution of present versus absent acute pain outcomes was a 50:50 split, the maximum top decile of lift for a perfect model would be 2. In comparison, if the ratio of a present versus absent acute pain outcome was 80:20, then the maximum theoretical lift for a perfect model would be 5. A value of one or less signals an inaccurate model (i.e., the percentage of patients with higher probability of pain as predicted by the model in a subset of the test set does not exceed the percentage of patients with pain in the full test set). The value reported in this work is the maximum cumulative lift.

All analyses were conducted using SAS Enterprise Miner 12.1 (SAS Institute, Cary, NC).

Results

A total of 8,071 subjects were included in this study, reflecting a convenience sample of patients available with pain scores on POD 1. A 5,031-patient subset of this sample also had documented pain scores on POD 3 due to continued hospitalization. For POD 1 outcomes, all 8,071 subjects were included. Table 1 provides an overview of the demographic and procedural characteristics of the POD 1 and 3 samples.

Pain Outcomes on POD 1 and 3

Of the 8071 subjects included in the POD 1 dataset, 4,267 (53%) reported suffering from moderate to severe pain on the first day after surgery (Table 2). For the POD 3 dataset, 2,256 (45%) reported suffering from moderate to severe pain on the third day after surgery, yielding

Table 2 Associations between POD 1 and 3 outcomes

			POD 3 Median Pain Score as Moderate to Severe	
			No	Yes
POD 1 median pain score = moderate to severe	No	3,804	No 2775	Yes 2256
	Yes	4,267	1,885 (37.5%)* 890 (17.7%)	470 (9.34%) 1,786 (35.5%)

* Percentages = Percentage of total POD 3 subject pool.

an absolute reduction rate of 8%. Of the 4267 patients who reported suffering from moderate to severe pain on POD1, 2,676 remained hospitalized on POD3, and 1,786 (79%) of these patients also reported moderate to severe postoperative pain on POD3. For those 3,804 patients with no reports of moderate to severe pain on POD1, 2,335 remained hospitalized on POD3, 1885 (81%) of whom also reported no episodes of moderate to severe pain on POD3.

Imputation of Missing Variables

The majority of missing value imputations were due to absence of "present on admission" flag data for the fifth (1,983 imputations), sixth (2,575), seventh (3,189), and eighth (3,779) listed comorbid conditions, followed by features pertaining to home medication use (1,402 imputations for each home medication) and the identities of the attending surgeon or anesthesiologist. A summary of the imputations for the POD1 and POD3 datasets can be found in Appendix Table A1.

Data Partition

As noted above, to avoid overfitting, algorithms were trained on the training set, tuned on the validation set, and then tested on the hold-out partition of data. For the POD1 data, there were 3,227 subjects partitioned to the training set, 2,421 to the validation set, and 2,423 to the hold-out data set. By design, 53% of patients suffered from moderate to severe pain in each of the three partitions. For the POD3 data, there were 2,011 subjects partitioned to the training set, 1,509 to the validation set, and 1,511 to the test set. Again by design, within the POD3 training set, 45% of patients suffered from moderate to severe pain in each of the three sets.

Feature Selection

Separate sets of features were selected for the POD 1 and 3 outcomes (Table 3). Details concerning patient age, type of surgery, and comorbidities grouped using the CCS array featured prominently in the POD 1 and 3 outcomes. Home opioid use carried a much higher relative importance for POD1 outcomes (relative importance 0.54) than POD3 (relative importance 0.26) outcomes.

Model Comparison

Each algorithm was compared on the hold-out test set using the full and reduced feature set, and then against the outcome of moderate to severe pain on POD 1 and 3, yielding a total of four experimental branches (Table 4). Overall, the LASSO algorithm, using the entire feature set to predict the occurrence of moderate to severe pain on POD 3, had the highest accuracy, with an area under the ROC of 0.727.

For POD 1, the LASSO algorithm, using the full feature set, had the highest accuracy with an ROC of 0.704. This was followed by the gradient-boosted decision tree algorithm, with an ROC of 0.665 and the k-NN algorithm, with an ROC of 0.643. In this branch of the experiment, the LASSO algorithm suffered 844 misclassifications for a misclassification rate of 0.35 (Figure 3A). Using the full feature dataset, the LASSO algorithm exhibited a cumulative lift of 1.49 given the 53% incidence of postoperative pain, suggesting that at the top decile, 78% of that decile's patients actually did suffer from severe acute postoperative pain (Figure 4A). On POD 1 using the full feature set, LASSO exhibited a sensitivity of 0.69, a specificity of 0.61, and a likelihood ratio of 1.77 (Table 5). Table 6 demonstrates those parameter estimates with the greatest weights when tested using the entire feature set via LASSO.

When using the full feature set on POD 1, the neural network and logistic regression algorithms had the lowest accuracy, with an ROC of 0.5 each. This suggested negligible improvement in classification accuracy over that offered by chance.

When the feature set was reduced using the prealgorithm variable selection step, the LASSO algorithm again had the highest accuracy, with an ROC of 0.704, followed by the gradient-boosted decision tree algorithm, with an ROC of 0.698, and the autoneural algorithm, with an ROC of 0.688. Here, accuracy of the LASSO algorithm remained grossly unchanged, committing 848 misclassifications versus 844 with the full feature set. However, the gradient-boosted decision tree algorithm had increased accuracy with the reduced feature set, increasing in ROC from 0.665 with the full feature set to

Table 3 Results of automated feature selection for pain prediction outcomes on POD 1 and 3

POD1			POD3		
Variable	Number of Rules in Tree	Relative Importance	Variable	Number of Rules in Tree	Relative Importance
Age	1	1	CCS code: Insertion, replacement, or removal of extracranial ventricular shunt	4	1
CCS group for secondary CPT code	5	0.88	Age	1	0.90
Home opioid	2	0.54	Admitting service	2	0.84
CCS diagnosis category 20: Cancer; other respiratory and intrathoracic	3	0.54	CCS diagnosis code 17: Cancer of Pancreas	4	0.82
Payor type	1	0.48	CCS diagnosis code 20: Cancer; other respiratory and intrathoracic	1	0.47
CCS procedure category 3: Laminectomy, excision of Intervertebral Disc	2	0.47	Surgical service	1	0.35
OR room name	2	0.37	CCS procedure code 4: Diagnostic spinal tap	1	0.32
Presence of more than 25 ICD9 comorbidities	1	0.36	CCS diagnosis code 18: Cancer of other GI organs; peritoneum	1	0.29
CCS diagnosis category 95: Other nervous system disorders	1	0.27	CCS procedure code 3: Laminectomy, excision of intervertebral disc	1	0.28
CCS diagnosis category 19: Cancer of bronchus; lung	1	0.23	Home opioid	1	0.26
CCS procedure code 5: Insertion of catheter or spinal stimulator into spinal canal	1	0.18	CCS diagnosis code 60: Acute posthemorrhagic anemia	1	0.26
CCS diagnosis code 98: Essential hypertension	1	0.17	CCS diagnosis code 35: Cancer of brain and nervous system	1	0.16
CCS diagnosis category 237: Complication of device; implant or graft	1	0.13	AHRQ code: Paralysis	1	0.16
BMI	1	0.13			
Patient status (inpatient vs. outpatient vs. patient on admission)	1	0.13			
Home SSRI/SSNRI	1	0.12			
Home benzodiazepine	1	0.09			

* Number of results in tree refers to how often this feature occurred in decision trees used for autonomous feature set selection. CCS diagnosis codes referred to the coding of one of up to 50 comorbidities within the CCS classification system.

0.698 with the reduced. Using the reduced-feature dataset, the LASSO algorithm exhibited a slightly lower cumulative lift of 1.44, suggesting that the top decile is 1.44 times more likely to include patients with severe acute postoperative pain than would a model based on random sampling.

Table 4 Comparison test outcomes of machine learning algorithms

Feature Set	POD	Algorithm	ROC	Misclassification Rate	Number of Wrong Classifications*	Cumulative Lift
Full feature set	1	LASSO	0.704	0.35	844	1.49
		Gradient boosting	0.665	0.38	916	1.47
		MBR	0.643	0.39	934	1.36
		SVM	0.627	0.40	975	1.37
		Autoneural	0.500	0.47	1,142	1.00
		Dmine regression	0.500	0.53	1,281	1.00
	3	LASSO	0.727	0.32	483	1.61
		Gradient boosting	0.682	0.38	573	1.58
		MBR	0.637	0.39	590	1.35
		SVM	0.635	0.40	604	1.29
		Autoneural	0.500	0.45	678	1.00
		Dmine regression	0.500	0.45	678	1.00
Reduced feature set	1	LASSO reduced	0.704	0.35	848	1.44
		Gradient boosting reduced	0.698	0.35	848	1.51
		Autoneural reduced	0.688	0.36	884	1.46
		Dmine regression reduced	0.628	0.40	975	1.37
		MBR reduced	0.601	0.44	1,055	1.28
		SVM reduced	0.592	0.46	1,104	1.26
	3	LASSO reduced	0.717	0.33	504	1.60
		Gradient boosting reduced	0.702	0.35	534	1.63
		Autoneural reduced	0.691	0.36	538	1.54
		SVM reduced	0.620	0.44	670	1.30
		Dmine regression reduced	0.599	0.41	623	0.87
		MBR reduced	0.539	0.50	751	1.18

* Number of wrong classifications is calculated from misclassification rate for gradient boosting decision tree.

As mentioned previously, for POD 3, the LASSO algorithm using the full feature set again had the highest accuracy, with a ROC of 0.727. In this branch of the experiment, the LASSO algorithm suffered 483 misclassifications for a misclassification rate of 0.32. This was again followed by the gradient-boosted decision tree, with an ROC of 0.682, and the k-NN algorithm, with a ROC of 0.637 (Figure 3B). Using the full feature dataset, the LASSO algorithm exhibited a cumulative lift of 1.61, suggesting that the top decile is 1.61 times more likely to include patients with severe acute postoperative pain than would a model based on random sampling (Figure 4B). On POD 3, with the full feature set, LASSO exhibited a sensitivity of 0.59, a specificity of 0.75, and likelihood ratio of 2.4.

When using the full feature set on POD 3, the neural network and logistic regression had the lowest accuracy, each with an ROC of 0.5. This suggested negligible improvement in classification accuracy over that offered by chance.

When the feature set was reduced using the prealgorithm variable selection step on POD 3 outcome data, the LASSO algorithm had the highest accuracy, with an ROC of 0.717, followed by the gradient-boosted decision tree algorithm, with an ROC of 0.702, and then neural network algorithm, with a ROC of 0.691. Using the reduced-feature dataset, the LASSO algorithm exhibited a cumulative lift of 1.6, suggesting that the LASSO algorithm detected, or captured, 61% of those subjects who suffered from moderate to severe postoperative pain.

Discussion

Our results demonstrate that machine learning algorithms, when applied to high-dimensional datasets developed from clinical data repositories, offer substantial improvements in accuracy over the tested logistic regression-based approaches to classification of acute postoperative pain outcomes. The majority of algorithms offered slightly better accuracy in predicting the occurrence of moderate to severe postoperative pain on POD 3 in comparison to POD 1. Reducing the number of predictor variables using used an automated approach improved the accuracy of many of the algorithms tested; however, LASSO performed equally well with the complete and reduced feature sets.

Our analysis included multiple metrics of algorithm performance to more fully delineate the differences in

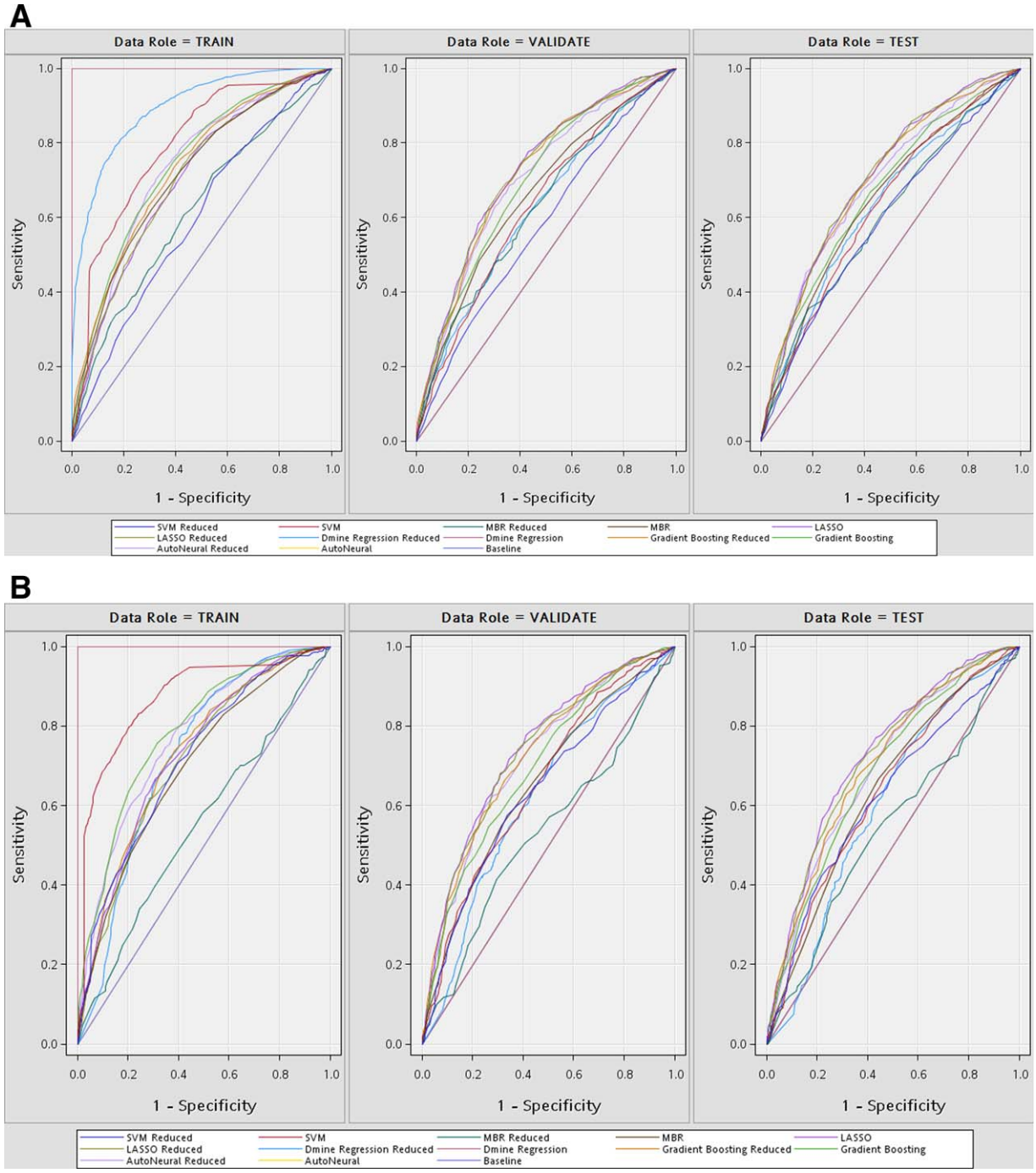

Figure 3 ROC for pain outcomes on POD 1 and 3. The ROC for each tested classifier are shown at the training, validation, and testing stages for POD 1 (A) and POD 3 (B). For POD 1, the LASSO algorithm, using the full feature set, had the highest accuracy, with a ROC of 0.704. For POD 3, the LASSO algorithm, using the full feature set, again had the highest accuracy, with a ROC of 0.727. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

prediction capabilities afforded by each machine learning approach. Although ROC is a widely accepted metric of model accuracy, it fails to provide substantial insight into what portion of the population is likely to benefit from the accuracy offered by the model [35]. This is partially due to the fact that the proportion of

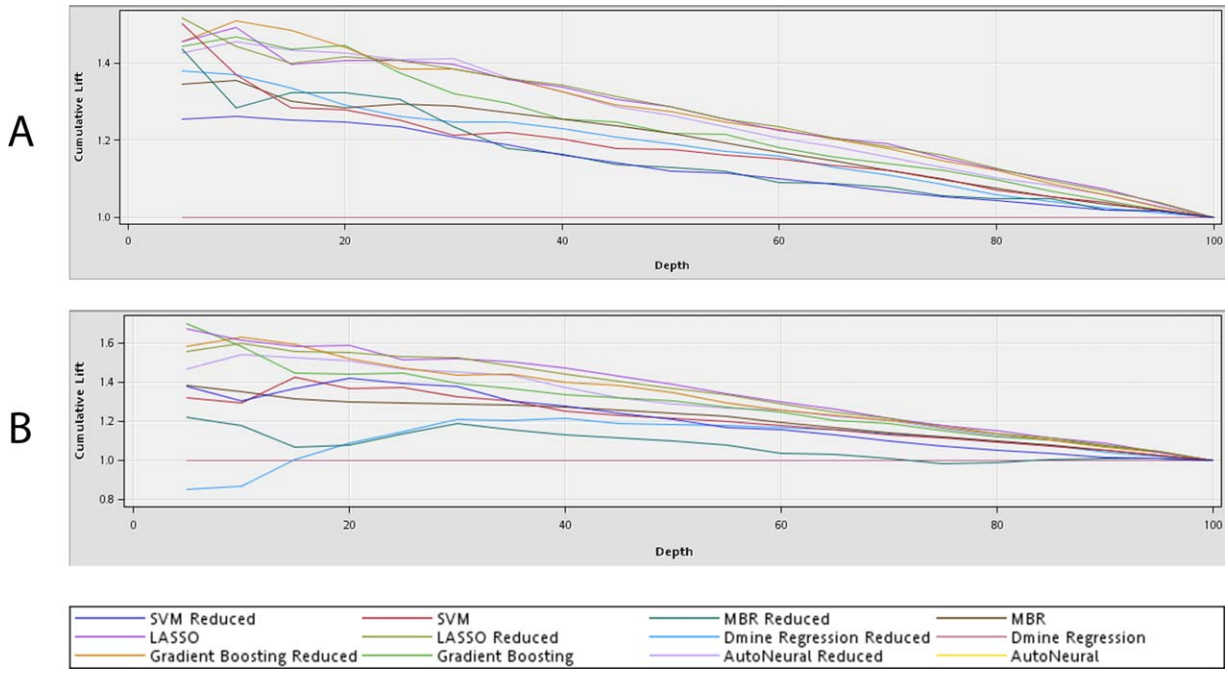

Figure 4 Cumulative lift curves for pain outcomes on POD 1 and 3. The LASSO algorithm exhibited a cumulative lift of 1.49 given the 53% incidence of moderate to severe postoperative pain on POD 1, suggesting that at the top decile, 78% of that decile's patients actually did suffer from severe acute postoperative pain. (A) On POD 3, the LASSO algorithm exhibited a cumulative lift of 1.61, suggesting the top decile is 1.61 times more likely to include patients with severe acute postoperative pain than would a model based on random sampling. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

patients who will suffer from moderate to severe postoperative pain is not equal to the proportion of those who will not. Subsequently, and even independently in some cases, misclassifications may be biased toward, or against, the detection of patients likely to suffer from moderate to severe pain after surgery. Indeed, the results presented here suggest that the LASSO algorithm may capture a larger proportion of patients expected to have an adverse acute pain outcome on POD1 than on POD3, despite having an identical ROC. This information may be helpful in developing future iterations of a postoperative pain prediction pipeline by modifying the costs associated with a particular misclassification, thereby helping influence the direction of misclassification to favor the detection of at-risk patients.

Using only routinely collected clinical data, our results compare favorably to the models derived from prior studies in which predictive models were prospectively developed using datasets designed a priori for research purposes [26,27]. Kalkman and others [27] prospectively examined 1,416 patients undergoing a mix of surgical procedures, excluding cardiac and neurosurgical cases, and developed a logistic regression model incorporating the following features: age, gender, type of surgery, intended incision size, blood pressure, heart rate, body mass index, preoperative pain intensity, and health-related quality of life as measured by the SF-36, the State-Trait Anxiety Inventory, and the Amsterdam Preoperative Anxiety and Information Scale. The bootstrapped model had an ROC of 0.73, and the authors concluded that pain scores within the first hour of surgery can be predicted using a set of variables collectible during a preoperative evaluation. Although this represented a significant contribution toward the prediction of postoperative pain, it should be noted that early postoperative pain scores do not correlate well with pain scores reported on POD 1 through 5 [39]. Furthermore, the work by Kalkman incorporated variables that were collected solely for the purpose of postoperative pain prediction; such tools are not universally applied in clinical preoperative evaluations. Our model accuracy of 0.7–0.73, using routinely available clinical data not prescreened for inclusion into the model, thus compares favorably to the dedicated prospective efforts by Kalkman.

Similarly, Sommer et al. collected postoperative pain scores on 1,490 patients undergoing a mix of surgical procedures [26]. Preoperative variables included

Table 5 Confusion matrix for LASSO with full feature set: POD 1 and 3

		Outcome: Moderate to Severe Pain	
POD1*	Yes/No	Yes	No
Prediction: moderate to severe pain	Yes	880	443
	No	401	699
POD3†			
		Yes	No
Prediction: moderate to severe pain	Yes	399	204
	No	279	629

* Misclassification rate = 0.348328518 False positive rate = 0.387915937
Sensitivity = 0.68696331 False negative rate = 0.31303669
Specificity = 0.612084063 Likelihood ratio = 1.770907675
PPV = 0.665154951 NPV = 0.635454545
† Misclassification rate = 0.319655857 False positive rate = 0.244897959
Sensitivity = 0.588495575 False negative rate = 0.411504425
Specificity = 0.755102041 Likelihood ratio = 2.403023599
PPV = 0.661691542 NPV = 0.692731278

demographics, type of anesthesia, type of surgery, American Society of Anesthesiologists score, duration of procedure, and multiple psychometric scales. ROC ranged from 0.74 on the day of surgery to 0.78 on POD 4, a trend similar to our results, suggesting an increase in model accuracy with each POD. Notably, there is no report of any type of validation step used by Sommer et al., raising the possibility that their results suffered from model overfitting. For comparison, our own results offered an ROC of 1, 0.89, and 0.79 for the unvalidated training-set models developed by the full feature set logistic regression, reduced feature set logistic regression, and full feature set SVM algorithms on POD 1.

Our pragmatic approach to postoperative pain prediction thus offers classification accuracy that, although less than ideal, compares quite favorably to prior published work. Moreover, while these prior groundbreaking reports are quite laudable in their scope and results, they nevertheless used approaches that lacked the ability to include additional variables. For instance, the inclusion of genomic data alone may result in the addition of tens of thousands of features for any given patient. Our approach suggests that pragmatic, autonomous forecasting of postoperative acute pain outcomes may be feasible for individual healthcare systems, thus permitting customization of models to the patients and practices that are particular to a given hospital and population.

Altogether, the risks and benefits associated with the assortment of pharmacologic and needle-based therapies offered by modern acute pain medicine services points to the need for accurate decision support systems capable of determining which patients are likely to benefit from such analgesic interventions. Simultaneously, such forecasts may spare those patients not at-risk for severe acute postoperative pain from the risks and costs inherent to regional anesthetics. Our results offer a specificity of 0.755 on POD 3, thus providing a moderate capability to spare those who would otherwise be scheduled for a nerve block from the associated risks and costs.

We also demonstrated a pragmatic application of advanced analytic methods to automatically process existing EMR data, select relevant variables, and then forecast severe acute postoperative pain [31]. The manual review of records to organize and "clean" data is no longer a feasible modeling approach given the massive amount of clinical data accumulated for each patient [40]. When using large administrative datasets, many patient characteristics that may be associated with poor postoperative pain outcomes, such as anxiety, catastrophizing, and socioeconomic status, may not be readily available in forms that are used within the experimental paradigm. Furthermore, the number of patients whose records would need to be reviewed in a time-sensitive fashion given the often short time interval between OR case scheduling and surgery makes this approach even more impractical. This presents a realistic challenge in converting experimentally derived models to models that are clinically applicable. This challenge, however, may be overcome with automated methods to processing EMR data, such as those presented here. Also, with the increasing structured clinical collection of social and behavioral characteristic, such as socioeconomic status, these automated methods may be made even more powerful in predicting postoperative pain.

Our data suffered from several limitations inherent to retrospective cohort studies. First, our study used static aggregate measures of pain by looking at the median pain scores. This represented a tradeoff in the specificity of the targeted outcome, such as would have been offered by selecting the number of severe pain events or focusing on severe pain events, for a more generalized clinical applicability affecting a larger proportion of patients. Second, this study did not incorporate information pertaining to analgesic use or functional capacity. Interestingly, we found that data pertaining to opioid administration via patient-controlled analgesia devices was not readily incorporated into the standard clinical EMR system. Although beyond the scope of this project (the use of machine learning classifiers), the simultaneous prediction of pain, analgesic requirement, and functional capacity remains an important goal for clinical decision support systems designed to forecast acute postoperative pain outcomes. Our model for POD 3 also suffered from a censuring effect, in that we offered no information pertaining to the reason for discharge of patients between POD 1 and 3. Discharges in this time

Table 6 Parameter estimates for LASSO on POD1 and 3

POD	Variable	Class Level	Standardized Estimate	Estimate
1	Intercept		0.0000	0.9591
	Age		−0.1794	−0.0057
	Home benzodiazepine	NO	−0.0042	−0.0062
	Home opioid	NO	−0.0682	−0.0688
	Home SSRI	NO	−0.0124	−0.0145
	Admitting surgical service	Trauma and acute care surgery	0.0190	0.0374
	Age Group	65–84	−0.0160	−0.0169
	CCS Category 205: Spondylosis	0	−0.0582	−0.0931
	CCS Category 212: Other bone disease and msk deformities	0	−0.0237	−0.0527
	CCS Procedure Category 3: Laminectomy, excision intervertebral disc	Yes	0.0037	0.0098
	Primary diagnosis of spondylosis	Yes	0.0266	0.0721
	Marital status	Divorced	0.0231	0.0399
	Primary CPT code category	MUSCULOSKELETAL	0.0356	0.0404
	Surgical service code	Cardiothoracic	−0.0557	−0.0872
	Surgical service code	Pancreas/hepatic/biliary	−0.0081	−0.0233
3	Age		−0.179	−0.006
	Home opioid	NO	−0.068	−0.069
	CCS Category 205: Spondylosis	0	−0.058	−0.093
	Surgical service code	Cardiothoracic	−0.056	−0.087
	Primary CPT code category	MUSCULOSKELETAL	0.036	0.040
	Primary diagnosis of spondylosis	Yes	0.027	0.072
	CCS Category 212: Other bone disease and msk deformities	0	−0.024	−0.053
	Marital status	Divorced	0.023	0.040
	Admitting surgical service	Trauma and acute care surgery	0.019	0.037
	Age group	65–84	−0.016	−0.017
	Home SSRI	NO	−0.012	−0.015
	Surgical service code	Pancreas/hepatic/biliary	−0.008	−0.023
	Home benzodiazepine	NO	−0.004	−0.006
	CCS procedure category 3: Laminectomy, excision intervertebral disc	Yes	0.004	0.010
	Intercept		0.000	0.959

interval may be due to low postoperative pain, whereas those patients remaining in the hospital may be there strictly due to poor pain control. This shortcoming points to the importance of supplemental data, as mentioned above, as well as the incorporation of time-domain information regarding resolution of acute pain, as explored preliminarily by Chapman et al. [41,42]. Perhaps the most important shortcoming of this study was the

overall lack of model accuracy demonstrated despite the use of advanced algorithms and a highly dimensional dataset. Our results compare favorably to those reported by Kalkman and Sommer et al., despite their inclusion of additional psychometric data selected to enhance prediction of postoperative pain, and despite the lack of validation of the model in one of the studies [26]. Nevertheless, a large proportion of the observed variance in postoperative pain outcomes remains unexplained by our model. Fortunately, the machine learning approach tested here is well positioned to incorporate even higher dimensional data, including genetic, text, and social network variables in future studies.

In summary, our results suggest the feasibility of an autonomous "analytic pipeline" as follows: on scheduling for surgery, the entire set of variables contained within a patient's EMR could be sent to a machine learning classification system that has previously been trained, validated, and tested using historical data from many patients who have recently undergone surgery in the health system. Next, the system would automatically clean the patient's data and forecast whether or not that patient is likely to suffer from moderate or severe pain after surgery. Those predictions could then be forwarded to the perioperative teams that would care for the patient on the day of surgery. Such an early-warning system may provide valuable information that allows a perioperative team to go beyond simple heuristics in choosing anesthesia therapies, such as basing them only on type of surgery. Notably, while an analytic pipeline based on the classification methods in this article would provide a clinically valuable prediction of pain risk, it would not suggest specific preventative, preemptive, or rescue analgesia for a given patient. However, future work could migrate our general analytic pipeline approach from simply forecasting postoperative pain to simultaneously considering the clinical context of the postoperative pain experience and recommending therapies.

Machine learning algorithms thus, when combined with highly dimensional datasets, offer an exciting opportunity to accurately forecast severe acute postoperative pain. Although our results demonstrate the feasibility with accuracy comparable to prior efforts, future work will need to improve the analyzed feature set as well the target pain-related outcomes.

References

1 Apfelbaum JL, Chen C, Mehta SS, Gan ATJ. Postoperative pain experience: Results from a National survey suggest postoperative pain continues to be undermanaged. Anesth Analg 2003;97:534–40.

2 Kehlet H, Jensen TS, Woolf CJ. Persistent postsurgical pain: Risk factors and prevention. Lancet 2006;367:1618–5.

3 Buvanendran A, Kroin JS. Multimodal analgesia for controlling acute postoperative pain. Curr Opin Anaesthesiol 2009;22:588–93.

4 Katz J, Clarke H, Seltzer Z. Preventive analgesia. Anesth Analg 2011;113:1242–53.

5 Ip HYV, Abrishami A, Peng PWH, Wong J, Chung F. Predictors of postoperative pain and analgesic consumption: A qualitative systematic review. Anesthesiology 2009;111:657–77.

6 Kalkman CJ, Visser K, Moen J, et al. Preoperative prediction of severe postoperative pain. Pain 2003; 105:415–23.

7 Sommer M, de Rijke JM, van Kleef M, et al. The prevalence of postoperative pain in a sample of 1490 surgical inpatients. Eur J Anaesthesiol 2008; 25:267–74.

8 Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be? Genome Biol 2013;14:205.

9 Okser S, Pahikkala T, Aittokallio T. Genetic variants and their interactions in disease risk prediction: Machine learning and network perspectives. BioData Mining 2013;6:1.

10 Menden MP, Iorio F, Garnett M, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One 2013;8:e61318.

11 Bessarabova M, Ishkin A, JeBailey L, Nikolskaya T, Nikolsky Y. Knowledge-based analysis of proteomics data. BMC Bioinf 2012;13:S13.

12 Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inf Assoc 2006;13:516–25.

13 DeLisle S, Kim B, Deepak J, et al. Using the electronic medical record to identify community-acquired pneumonia: Toward a replicable automated strategy. PLoS One 2013;8:e70944.

14 Phil Simon. Too Big to Ignore: The Business Case for Big Data—Google Books 2013. Available at: http://books.google.com/books?hl=en&lr=&id=DnGdoh66sgC&oi=fnd&pg=PR3&dq=too+big+to+ignore+the+business+case+for+big+data&ots=VH0wrZWms6&sig=bP45pjvVUqqOuzHXWzFgNDtXEpM.

15 Breiman L. Statistical modeling: The two cultures. Stat Sci 2001. doi:10.1006/aama.1996.0501.

16 Hall M, Franke E, Holmes G, et al. The WEKA data mining software: An update. SIGKDD Explor 2009;11:10–8.

17 Witten I, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition. San Francisco: Morgan Kaufmann, 2005.

18 Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935–42.

19 Li J, Huang K-Y, Jin J, Shi J. A survey on statistical methods for health care fraud detection. Health Care Manage Sci 2008;11:275–87.

20 Bolton RJ, Hand DJ. Statistical fraud detection: A review. Stat Sci 2002. doi:10.2307/3182781.

21 Labusch K, Barth E, Martinetz T. Simple method for high-performance digit recognition based on sparse coding. IEEE Trans Neural Netw 19:1985–9.

22 Zorkadis V, Karras DA, Panayotou M. Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering. Neural Netw 2005;18:799–807.

23 Furukawa MF. Meaningful use: A roadmap for the advancement of health information exchange. Isr J Health Policy Res 2013;2:1.

24 Lai M, Kheterpal S. Creating a real return-on-investment for information system implementation: Life after HITECH. Anesthesiol Clin 2011;29:413–38.

25 Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010;363:501–4.

26 Sommer M, de Rijke JM, van Kleef M, et al. Predictors of acute postoperative pain after elective surgery. Clin J Pain 2010;26:87–94.

27 Kalkman CJ, Visser K, Moen J, et al. Preoperative prediction of severe postoperative pain. Pain 2003; 105:415–23.

28 Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J Chronic Dis 1987;40:373–83.

29 Cohen PR, Jensen D. Overfitting explained. In: Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics. 1997: 115–22.

30 Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. J Clin Epidemiol 2008;61:1085–94.

31 Babyak MA. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004; 66:411–21.

32 Tao KM. A closer look at the radial basis function (RBF) networks. Neurocomputing 1997;14:273–88.

33 Lee S-M, Abbott P, Johantgen M. Logistic regression and Bayesian networks to study outcomes using large data sets. Nursing Res 2005;54:133–8.

34 Statistical pattern recognition: A review. IEEE Trans Pattern Anal Mach Intell 2000;22(1):4–37.

35 Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–98.

36 Linden A. Measuring diagnostic and predictive accuracy in disease management: An introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract 2006;12:132–9.

37 Provost FJ, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. ICML 1998.

38 Bhattacharyya S. Evolutionary algorithms in data mining: Multi-objective performance modeling for direct marketing. In: Proceedings Sixth ACM SIGKDD International Conference Knowledge Discovery Data Mining, New York, 2000: 465–73.

39 Tighe PJ, Harle CA, Boezaart AP, Aytug H, Fillingim R. Of rough starts and smooth finishes: Correlations between post-anesthesia care unit and postoperative days 1–5 pain scores. Pain Med 2014;15:306–15.

40 Zurada J, Lonial S. Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry. J Appl Business Res 2005:21.

41 Chapman CR, Donaldson GW, Davis JJ, Bradshaw DH. Improving individual measurement of postoperative pain: The pain trajectory. J Pain 2011;12:257–62.

42 Chapman CR, Donaldson G, Davis J, Ericson D, Billharz J. Postoperative pain patterns in chronic pain patients: A pilot study. Pain Med 2009;10:481–7.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix

Table A1 Summary of missing value imputations

Dataset	Variable	Number of Missing Value Imputations
POD1	Attending anesthesiologist	157
	Attending surgeon	334
	Home benzodiazepine	1,402
	CPT code #7	1
	Circulator RN	2
	Home muscle relaxant	1,402
	Home NSAID	1,402
	Home opioid	1,402
	POA1	2
	POA2	331
	POA3	801
	POA4	1,379
	POA5	1,983
	POA6	2,575
	POA7	3,189
	POA8	3,779
	Patient admission status	15
	Timing of surgery	957
	Home SSRI/SSNRI	1,402
	Actual surgical service	1
	Surgical service	1
POD3	Attending anesthesiologist	88
	Attending surgeon	230
	Benzo	855
	CPT Code #7	1
	Circulator RN	1
	Home muscle relaxant	855
	Home NSAID	855
	Home opioid	855
	POA1	2
	POA2	91
	POA3	274
	POA4	533
	POA5	833
	POA6	1,145
	POA7	1,509
	POA8	1,863
	POA9	2,194
	Patient admission status	10
	Timing of surgery	599
	Home SSRI/SSNRI	855

POA = Present on admission flag.