


RESEARCH

Open Access



# Exploring supervised machine learning approaches to predicting Veterans Health Administration chiropractic service utilization

Brian C. Coleman<sup>1,2\*</sup> , Samah Fodeh<sup>1,2</sup>, Anthony J. Lisi<sup>1,2</sup>, Joseph L. Goulet<sup>1,2</sup>, Kelsey L. Corcoran<sup>1,2</sup>, Harini Bathulapalli<sup>1,2</sup> and Cynthia A. Brandt<sup>1,2</sup>

## Abstract

**Background:** Chronic spinal pain conditions affect millions of US adults and carry a high healthcare cost burden, both direct and indirect. Conservative interventions for spinal pain conditions, including chiropractic care, have been associated with lower healthcare costs and improvements in pain status in different clinical populations, including veterans. Little is currently known about predicting healthcare service utilization in the domain of conservative interventions for spinal pain conditions, including the frequency of use of chiropractic services. The purpose of this retrospective cohort study was to explore the use of supervised machine learning approaches to predicting one-year chiropractic service utilization by veterans receiving VA chiropractic care.

**Methods:** We included 19,946 veterans who entered the Musculoskeletal Diagnosis Cohort between October 1, 2003 and September 30, 2013 and utilized VA chiropractic services within one year of cohort entry. The primary outcome was one-year chiropractic service utilization following index chiropractic visit, split into quartiles represented by the following classes: 1 visit, 2 to 3 visits, 4 to 6 visits, and 7 or greater visits. We compared the performance of four multiclass classification algorithms (gradient boosted classifier, stochastic gradient descent classifier, support vector classifier, and artificial neural network) in predicting visit quartile using 158 sociodemographic and clinical features.

**Results:** The selected algorithms demonstrated poor prediction capabilities. Subset accuracy was 42.1% for the gradient boosted classifier, 38.6% for the stochastic gradient descent classifier, 41.4% for the support vector classifier, and 40.3% for the artificial neural network. The micro-averaged area under the precision-recall curve for each one-versus-rest classifier was 0.43 for the gradient boosted classifier, 0.38 for the stochastic gradient descent classifier, 0.43 for the support vector classifier, and 0.42 for the artificial neural network. Performance of each model yielded only a small positive shift in prediction probability (approximately 15%) compared to naïve classification.

(Continued on next page)

\* Correspondence: [Brian.Coleman2@VA.gov](mailto:Brian.Coleman2@VA.gov)

<sup>1</sup>Pain Research, Informatics, Multimorbidities, and Education (PRIME) Center, VA Connecticut Healthcare System, 11-ACSL-G, 950 Campbell Avenue, West Haven, CT 06516, USA

<sup>2</sup>Yale School of Medicine, Yale University, New Haven, CT, USA



© US Government 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

**Conclusions:** Using supervised machine learning to predict chiropractic service utilization remains challenging, with only a small shift in predictive probability over naïve classification and limited clinical utility. Future work should examine mechanisms to improve model performance.

**Keywords:** Machine learning, Predictive Modeling, Chiropractic, Healthcare service utilization

## Introduction

### Rationale

Chronic pain is highly prevalent and carries a high cost burden, conservatively estimated at over \$560 billion annually and exceeding that of cardiovascular, neoplastic, and metabolic diseases [1]. In 2016, an estimated 20.4% of U.S. adults (50.0 million) experienced chronic pain and 8.0% (19.6 million) experienced high-impact chronic pain, with a higher prevalence in adults receiving public healthcare coverage [2].

Spinal pain conditions, including low back pain and neck pain, are among the most common musculoskeletal pain conditions and contribute greatly to the high prevalence of chronic pain, with 6.0% of U.S. adults experiencing chronic spinal pain and 2.2% experiencing high-impact chronic spinal pain [3]. Spinal pain conditions carry high spine-related and overall healthcare costs, with an average of \$3915 spent on spine-related healthcare and \$9781 on overall healthcare costs per chronic spinal pain patient per year [3]. High-impact chronic spinal pain carries even greater direct costs (\$5979 for spine-related healthcare and \$14,661 for overall healthcare). Indirect costs, including lost productivity due to disability, are also exceptionally high in this population [4].

Conservative interventions for spinal pain conditions have been associated with lower healthcare costs and improvements in pain status in different clinical populations, including veterans [5–9]. Veterans with musculoskeletal pain conditions, especially those of chronic nature, often utilize non-pharmacological pain management recommended by clinical practice guidelines, including U.S. Department of Veterans Affairs (VA) chiropractic care [10–13]. Many veterans with musculoskeletal pain conditions, including those receiving chiropractic care, demonstrate high prevalence of comorbid medical and mental health conditions [14–16], with those having higher comorbidity burdens demonstrating greater healthcare utilization [17].

Little is currently known about predicting healthcare service utilization in the domain of conservative intervention for spinal pain conditions. The proliferation of data available in the electronic health record (EHR) has led to a growing demand for prospective data-driven decision making through predictive analytics. Massive quantities of patient data are now easily accessible and rapidly queryable to enhance medical decision making,

support data visualization, and create data repositories that can be used to develop predictive models [18]. The large scale of routinely collected data support the utility of predictive models compared to traditional clinical prediction rules that require parsimonious criteria, easy computability, and independent validation that may take years to complete [19]. Predictive models have demonstrated great utility and potential in many clinical disciplines where prospective prediction can inform patient management and outcomes, aid in system-level resource allocation and logistics, and afford the opportunity for cost containment [19, 20]. Data collected in EHRs can be preprocessed, mined, and subsequently support real-time, point-of-care decision making through automated processes that enable scalability across systems and clinical disciplines [21].

Predictive analytics relevant to chiropractic practice has been limited to clinical prediction rules associated with response to spinal manipulation [22, 23]. Prior studies have examined components of service utilization as “dose” and “frequency” effects of spinal manipulation, however definitions of these terms vary considerably across studies [24]. Visit frequency recommendations included in clinical practice guidelines have been largely based on Delphi panels of expert opinions [25, 26], with current evidence suggesting that spinal manipulative treatment visit frequency does not significantly impact clinical outcomes during and following the treatment period [24].

Studies have not yet examined prediction of chiropractic service utilization. VA provides an important setting to examine chiropractic service utilization as the largest integrated healthcare system in the United States with an enterprise-wide health information system supporting system-level examination of comprehensive EHR data [27]. As a capitated delivery model, VA EHR data also affords the ability to examine chiropractic service utilization as relatively independent of third-party reimbursement influence, compared to traditional delivery settings where the fee-for-service structure may confound utilization.

### Objective

Effectively predicting healthcare service utilization may help to improve care delivery and inform resource allocation. In this proof-of-concept work, we aim to explore

the utility of a supervised machine learning approach in predicting one-year chiropractic service utilization by veterans receiving VA chiropractic care.

## Methods

The predictive models in this study were developed and reported in accordance with published recommendations for reporting machine learning models [28]. This study was approved by the Institutional Review Board at the VA Connecticut Healthcare System.

### Setting and dataset

The Musculoskeletal Diagnosis (MSD) Cohort, a cohort study using comprehensive national EHR data to examine musculoskeletal pain and pain care of veterans, was used as the data source for this study [14]. To be included in the MSD cohort, a veteran had to have one of 1685 International Classification of Diseases, 9th revision (ICD-9) musculoskeletal disorder diagnoses. Diagnoses had to be recorded during two or more outpatient visits within 18 months or during at least one inpatient stay. Additional sociodemographic and clinical data were extracted from the VA Corporate Data Warehouse for eligible veterans to allow for longitudinal analyses following entry into the cohort.

Figure 1 summarizes the collection, processing, and flow of data through our study. For this study, we included veterans who entered the MSD Cohort between October 1, 2003 and September 30, 2013 with at least one visit to on-station VA chiropractic services (denoted by the VA clinic stop code “436”) within 365 days of entering the MSD Cohort. This was done to ensure demographic and clinical data collected at the veteran’s MSD Cohort entry was reasonably proximate to the veteran’s first (index) chiropractic visit. One-year chiropractic service utilization was examined as the total visit frequency obtained by counting the number of visits over a period of 365 days following the veteran’s index chiropractic visit, with associated ICD-9 diagnosis codes obtained for each visit.

The diagnosis category of the index chiropractic visit was included in the final dataset. Visits were categorized as “Low back pain only”, “Neck pain only”, “Both low back and neck pain”, or “Neither low back nor neck pain” using an existing framework for identifying back and neck pain disorders in administrative data based on ICD-9 diagnoses [29].

Additional sociodemographic and clinical data were obtained for each veteran from their EHR, including cohort entry date, index chiropractic visit date, age at index chiropractic visit, index chiropractic visit facility, gender, period of service, service connected disability status, marital status, pain intensity numerical rating scale (NRS) score, body mass index (BMI), race/

ethnicity, smoking status, and Charlson Comorbidity Index (CCI). Additional clinical data included binary classification of the presence of many medical comorbidities, mental health comorbidities, musculoskeletal comorbidities, and prescription data within the veteran’s VA medical record.

Statistical analyses of clinical and sociodemographic variables across independent groups used single factor ANOVA for continuous variables and chi-square tests for categorical variables, with a significance level of 0.05.

### Prediction problem

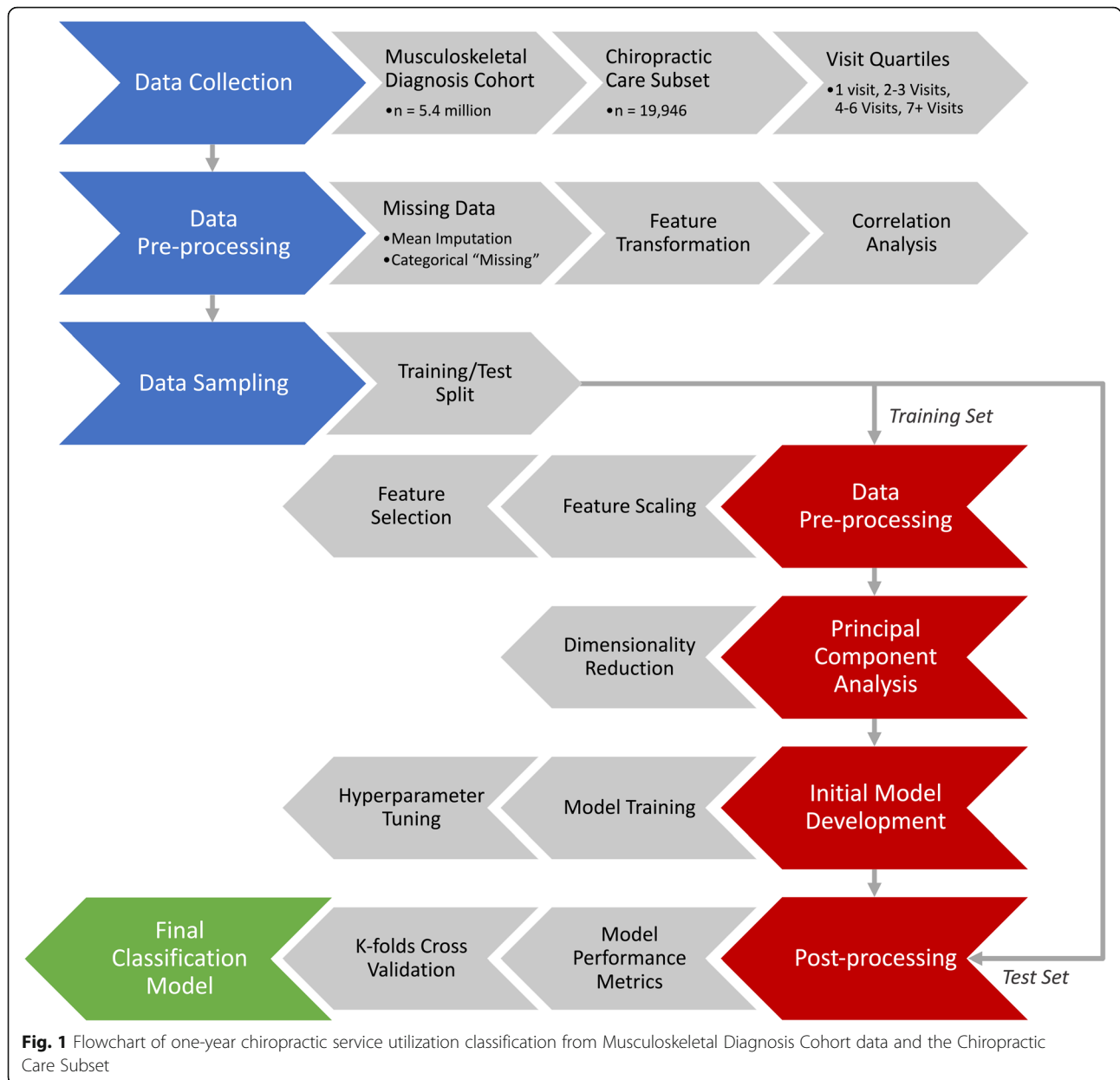
We sought to predict one-year chiropractic service utilization as a retrospective, prognostic multiclass classification problem. One-year chiropractic service utilization was categorized for each veteran into quartiles as the class label, determined by the distribution of the entire dataset. This was done to create more uniformly distributed classes, to protect against the influence of outliers showing high service utilization skewing the model, and to allow a single label, multiclass classification approach using one-versus-rest classification.

### Data preparation, feature selection, and feature engineering

Additional preprocessing was performed on the included features in the dataset. Entry year into the MSD Cohort was transformed as the number of years since 2003, the minimum entry year in this sample. The MSD Cohort entry date and the index chiropractic visit date were transformed such that the difference between the two dates (in days) was included as an engineered feature. All categorical variables were binarized into unique features using one-hot encoding. Mean imputation was used for cases with missing continuous body mass index data ( $n = 278$ , 1.4%). Missing categorical data were not imputed, with “Unknown” and “Unknown/Missing” as valid data entries for marital status, smoking status, and pain status.

No features demonstrated a clinical relationship or a strong correlation to one-year chiropractic service utilization, with the potential for information leakage limited. To account for collinear relationships between independent features affecting the model performance, features demonstrating a strong correlation with other features, using a Pearson correlation coefficient greater than 0.7 as the cutoff point, were dropped from the final dataset [30]. The final dataset included 158 features (Additional file 1).

For the initial phase of model development, the dataset was randomly partitioned with 70% allocated to the training set and 30% to the test set. Feature scaling was done using a standard scaler (with zero-mean and unit-variance), trained only on the training set and applied to



the training and test sets. Feature selection using a filter technique based on chi-squared analyses was performed to evaluate the impact of selecting a subset of features on prediction accuracy [31].

**Principal component analysis**

Principal component analysis (PCA) can be used to transform multi-dimensional data into fewer dimensions by geometrically projecting them into summary components that represent the variability, patterns, and relationships of the original features [32]. We used PCA on the scaled training set to reduce the dimensionality from 158 features into two principal components to visualize the separability of the four classes in two-dimensional

space. We also explored whether PCA may be useful to reduce the dimensionality of our dataset into principal components representing the variability of our data and the predicted label to improve our classification performance. We used the proportion of total variance explained to determine how many principal components to retain, which has been recommended in exploratory analyses and establishes a predetermined threshold of proportion of total variance explained (often between 70 to 95%) [33]. We examined the total number of principal components required to reach a threshold of 70 and 95%, the upper and lower bounds of the recommended range, before determining whether to proceed with using PCA as inputs to our model.

### Selecting and building the model

As an exploratory study, we selected a subset of available multiclass one-versus-rest classifiers to evaluate performance, based on preliminary sweeping of available classifiers using Python 3.5 and the SciKit-Learn (Version 0.19.0) library [34]. Four models were selected: a gradient boosted classifier, a stochastic gradient descent classifier, a linear support vector classifier, and an artificial neural network. A description of each selected model, including details on the hyperparameters and architectural parameters used in this study, is available as Additional File 2. Support-weighted precision, recall, F-measure, and subset accuracy (the percentage of total number of labels correctly predicted) were obtained for each algorithm in the initial development phase, with hyperparameters determined by grid-search and trial and error to maximize F-measure.

Using a one-versus-rest classification approach for each class (visit quartile) created four separate binary classifiers fitting one class against all other classes for each selected model. For the gradient boosted and stochastic gradient descent classifiers, the “OneVsRestClassifier” function of SciKit-Learn was used to build the series of binary classifiers. For the linear support vector classifier and multi-layered perceptron neural network, the additional function was not necessary as each has inherent multiclass capabilities using a one-versus-rest approach. As each class was represented by a single classifier (for each algorithm), this approach provided the advantage of being able to examine performance of the estimator for each class. A Precision-Recall curve (PRC) plot and the area under the PRC curve (AUC) were obtained for the binarized output of each one-versus-rest classifier to compare and evaluate performance of each algorithm for each class. The PRC was preferred to the Receiver Operating Characteristic curve, given the imbalance of predicting one quartile of the data against the remaining three [35].

Following initial development, we performed 10 cycles of 10-fold repeated, stratified cross-validation, for a total of 100 validation performances, to evaluate performance of the developed models. Feature scaling was done using a standard scaler during each validation iteration. Support-weighted precision, recall, F-measure, and subset accuracy were obtained for each iteration to compare performance metrics for each algorithm.

## Results

### Final model and performance

There were 19,946 veterans across 38 VA facilities who entered the MSD Cohort between October 1, 2003 and September 30, 2013 and had at least one visit to on-station VA chiropractic services within 1 year of MSD Cohort entry. Veteran sociodemographic and clinical characteristics are presented in Table 1. One-year

chiropractic service utilization ranged from 1 visit to 73 visits and was split into quartiles representing the following classes: 1 visit, 2 to 3 visits, 4 to 6 visits, and 7 or greater visits. The distribution of classes was nearly balanced, with the first and second quartiles slightly larger than 25% of the entire dataset. Feature selection using chi-squared analyses did not affect the classifier prediction accuracy. There was poor separability of all four classes of the target variable with reduction to two-dimensions using PCA (Fig. 2). The peak explained variance ratio for an individual principal component was 2.9%, with 80 principal components required to reach a proportion of total variance explained of 70% and 126 needed to reach 95%. As such, we chose to retain all 158 features without reducing dimensionality for the purpose of predicting one-year chiropractic service utilization. We felt our sample size was sufficient to handle retaining all 158 features, given a minimum events per variable ratio of over 27:1 in predicting the smallest class.

Performance metrics by class for each algorithm in the initial development phase are presented in Table 2. Overall performance of the four models was poor, with no model able to predict one-year chiropractic service utilization with a support-weighted subset accuracy greater than 42.1%. Precision, recall, and F-measure were similarly generally poor across all classes in all models during initial development.

The PRC and AUC for each one-versus-rest classifier in each model is presented in Fig. 3. Each classifier was better at identifying service utilization in the first or fourth quartile (1 visit or 7 or greater visits) than those within in the interquartile range. The range of the AUC for the micro-averaged PRC was 0.38 to 0.43. Given the baseline probability of a positive outcome in each one-versus-rest classifier of approximately 25%, this represents a small positive shift in prediction probability (approximately 15%) compared to naïve classification.

In the cross-validation phase, we found similar performance results (Fig. 4). The gradient boosted classifier, the support vector classifier, and the artificial neural network performed most consistently (median accuracy 41.5%, 41.1, and 39.7%, respectively). The stochastic gradient descent classifier performed most inconsistently, with the largest tails. The mean precision (with 95% confidence interval) was  $39.4 \pm 0.3\%$  for the gradient boosted classifier,  $24.8 \pm 2.0\%$  for the stochastic gradient descent classifier,  $38.7 \pm 0.3\%$  for the support vector classifier, and  $34.5 \pm 0.9\%$  for the artificial neural network. The mean recall (with 95% confidence interval) was  $41.5 \pm 0.2\%$  for the gradient boosted classifier,  $26.8 \pm 0.5\%$  for the stochastic gradient descent classifier,  $41.1 \pm 0.2\%$  for the support vector classifier, and  $39.7 \pm 0.2\%$  for the artificial neural network. The mean F-measure (with 95% confidence interval) was  $38.1 \pm 0.2\%$  for the gradient

**Table 1** Patient sociodemographic and clinical characteristics

Variable	Total	Visits within 1 year				p Value
		1 Visit	2–3 Visits	4–6 Visits	7+ Visits	
N	19,946	5473 (27.4)	5233 (26.2)	4280 (21.5)	4960 (24.9)	
Age, median [IQR], y	45 [30–58]	44 [29–57]	44 [30–58]	46 [30–59]	47 [34–59]	< .00001
Sex						
Female	13.5	25.6	24.6	21.9	27.9	< .001
Male	86.5	27.7	26.5	21.4	24.4	
Index Chiropractic Visit Diagnosis						< .00001
Low Back Pain Only	55.6	30.5	26.3	21.1	22.2	
Neck Pain Only	9.0	27.6	26.3	21.6	24.5	
Both Low Back and Neck Pain	31.3	20.5	26.5	22.2	30.8	
Neither Low Back nor Neck Pain	4.1	39.1	24.0	20.4	16.4	
Race						< .00001
White	70.7	26.8	26.1	21.4	25.7	
Black	12.2	28.0	25.1	21.8	25.2	
Hispanic	7.7	27.0	27.8	22.7	22.5	
Other	2.4	26.7	23.6	24.4	25.4	
Unknown	7.1	33.6	28.5	19.0	18.9	
Pain intensity, median [IQR] <sup>a</sup>	4 [0–6]	4 [0–6]	4 [0–6]	4 [0–6]	4 [1–7]	0.057
No Pain or Mild Pain Intensity (NRS 0–3)	44.3	28.8	26.6	21.2	23.4	0.190
Moderate to Severe Pain Intensity (NRS 4–10)	55.7	28.0	26.0	21.2	24.8	
Smoking Status <sup>b</sup>						< .00001
Never	36.3	27.4	25.4	21.5	25.7	
Former	39.9	27.9	27.8	21.3	23.0	
Current	23.8	23.8	25.8	22.4	28.0	
BMI, mean (SD), kg/m <sup>2</sup> <sup>c</sup>	29.4 (5.4)	29.1 (5.2)	29.1 (5.4)	29.1 (5.4)	29.3 (5.4)	0.054
Not obese, BMI < 30 kg/m <sup>2</sup>	61.6	27.5	26.1	21.8	24.6	0.311
Obese, BMI ≥ 30 kg/m <sup>2</sup>	38.4	27.1	26.4	20.9	25.5	
Period of Service						< .00001
OEF/OIF/OND	29.7	28.9	28.1	21.1	21.9	
Gulf War	24.3	29.1	25.7	20.8	24.4	
Post-Vietnam Era	12.8	25.9	25.2	22.2	26.7	
Vietnam	26.9	25.9	25.1	21.7	27.3	
Other	6.4	23.5	26.5	23.3	26.7	
Marital Status						< .01
Married	49.1	27.6	25.9	21.2	25.4	
Not Married	19.5	28.3	26.8	21.9	23.0	
Separated/Divorced	28.5	26.5	26.6	21.8	25.1	
Widow/Widower	2.3	25.3	25.8	20.9	28.0	
Unknown	0.5	44.3	23.6	12.3	19.8	
Service Connected Disability	65.4	26.9	25.7	21.4	26.0	< .00001
CCI, mean (SD)	0.40 (0.98)	0.36 (0.95)	0.39 (0.98)	0.40 (0.96)	0.43 (1.00)	< .01
CCI = 0	78.0	28.1	26.3	21.3	24.3	< .001
CCI ≥ 1	22.0	25.2	25.9	21.9	27.0	

**Table 1** Patient sociodemographic and clinical characteristics (Continued)

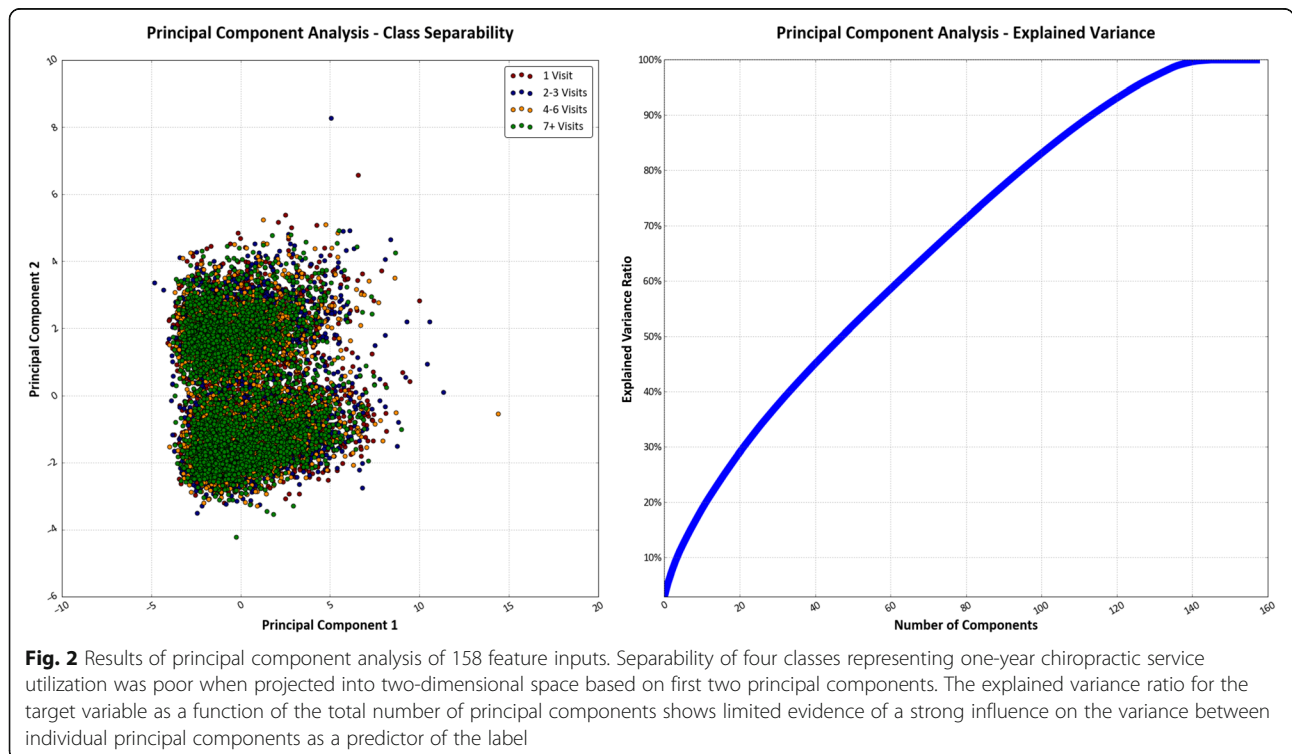
Variable	Total	Visits within 1 year				p Value
		1 Visit	2–3 Visits	4–6 Visits	7+ Visits	
Pharmaceutical Use <sup>d</sup>						
Opioid Prescription	13.3	28.3	26.3	19.1	26.4	.010
Tramadol Prescription	8.0	27.8	26.9	19.5	25.8	.238
Medical Comorbidities						
PTSD	20.1	27.3	26.5	21.5	24.7	.967
Mild Depression	21.3	26.2	25.9	20.5	27.5	< .0001
Major Depression	7.9	24.7	26.1	22.8	26.4	.051
Schizophrenia	0.5	27.9	29.8	15.4	26.9	.477
Bipolar	4.4	24.7	28.8	21.4	25.1	.199
TBI	4.7	29.4	28.2	21.9	20.6	.017
Alcohol or Substance Use Disorder	10.7	27.1	27.4	21.0	24.6	.663

BMI Body mass index; IQR Interquartile range; NRS Numerical rating scale; CCI Charlson Comorbidity Index; PTSD Post-traumatic stress disorder; TBI Traumatic brain injury; Significance at  $\alpha = 0.05$ ; <sup>a</sup> 3621 other/missing; <sup>b</sup> 685 other/missing; <sup>c</sup> 278 missing; <sup>d</sup> Prescription within 30 days of MSD Cohort entry

boosted classifier,  $16.9 \pm 1.0\%$  for the stochastic gradient descent classifier,  $37.3 \pm 0.2\%$  for the support vector classifier, and  $34.6 \pm 0.4\%$  for the artificial neural network. The mean accuracy (with 95% confidence interval) was  $41.5 \pm 0.2\%$  for the gradient boosted classifier,  $26.8 \pm 0.5\%$  for the stochastic gradient descent classifier,  $41.1 \pm 0.2\%$  for the support vector classifier, and  $39.7 \pm 0.2\%$  for the artificial neural network.

**Discussion**

Effectively predicting healthcare service utilization has multiple clinical implications and can help to improve delivery, population health, and resource allocation to support the Quadruple Aim and support transitions towards value-based care delivery systems [36, 37]. Predictive models have previously been developed and validated to predict healthcare resource utilization using



**Table 2** Classification matrix and subset accuracy of machine learning models to predict one-year chiropractic service utilization, based on parameters from initial development phase

Model/Class	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
<b>Gradient Boosted Classifier</b>				
1 Visit	43.5	61.7	51.0	
2–3 Visits	36.3	32.5	34.3	
4–6 Visits	34.0	9.2	14.5	
7+ Visits	46.2	57.8	51.0	
Average <sup>a</sup>	40.3	42.1	39.1	42.1
<b>Stochastic Gradient Descent Classifier</b>				
1 Visit	43.9	53.1	48.1	
2–3 Visits	32.3	29.9	31.0	
4–6 Visits	24.9	14.8	18.6	
7+ Visits	43.4	51.2	47.0	
Average <sup>a</sup>	36.8	38.6	37.2	38.6
<b>Support Vector Classifier</b>				
1 Visit	42.6	60.8	50.1	
2–3 Visits	35.3	26.1	30.0	
4–6 Visits	32.0	11.3	16.7	
7+ Visits	45.3	60.3	51.7	
Average <sup>a</sup>	39.2	41.4	38.4	41.4
<b>Artificial Neural Network</b>				
1 Visit	42.9	56.2	48.7	
2–3 Visits	35.9	30.6	33.1	
4–6 Visits	25.3	12.1	16.4	
7+ Visits	45.1	55.9	49.9	
Average <sup>a</sup>	38.0	40.3	38.2	40.3

<sup>a</sup>Support-weighted average

patient-level EHR data [37–39]. Further, they have demonstrated outperformance of existing clinical prediction rules, with machine learning models demonstrating small performance benefits (with limited clinical differences) over statistical models [38].

Current evidence to predict the use of chiropractic services over a discrete period of time is limited. In this study, we developed four machine learning models using a large cohort of veterans receiving VA chiropractic services. While we establish a baseline by which future models may be evaluated as a proof-of-concept work, the clinical utility of these models, based on the data used in this study, may be limited at this time.

Our models yield a small positive shift in predictive probability over naïve classification but remain limited by the high amount of false positives and false negatives. Given the average precision (positive predictive value), less than half of those that are predicted in a certain class are truly in that class, resulting in many false positives. Based on the average recall (sensitivity), less than

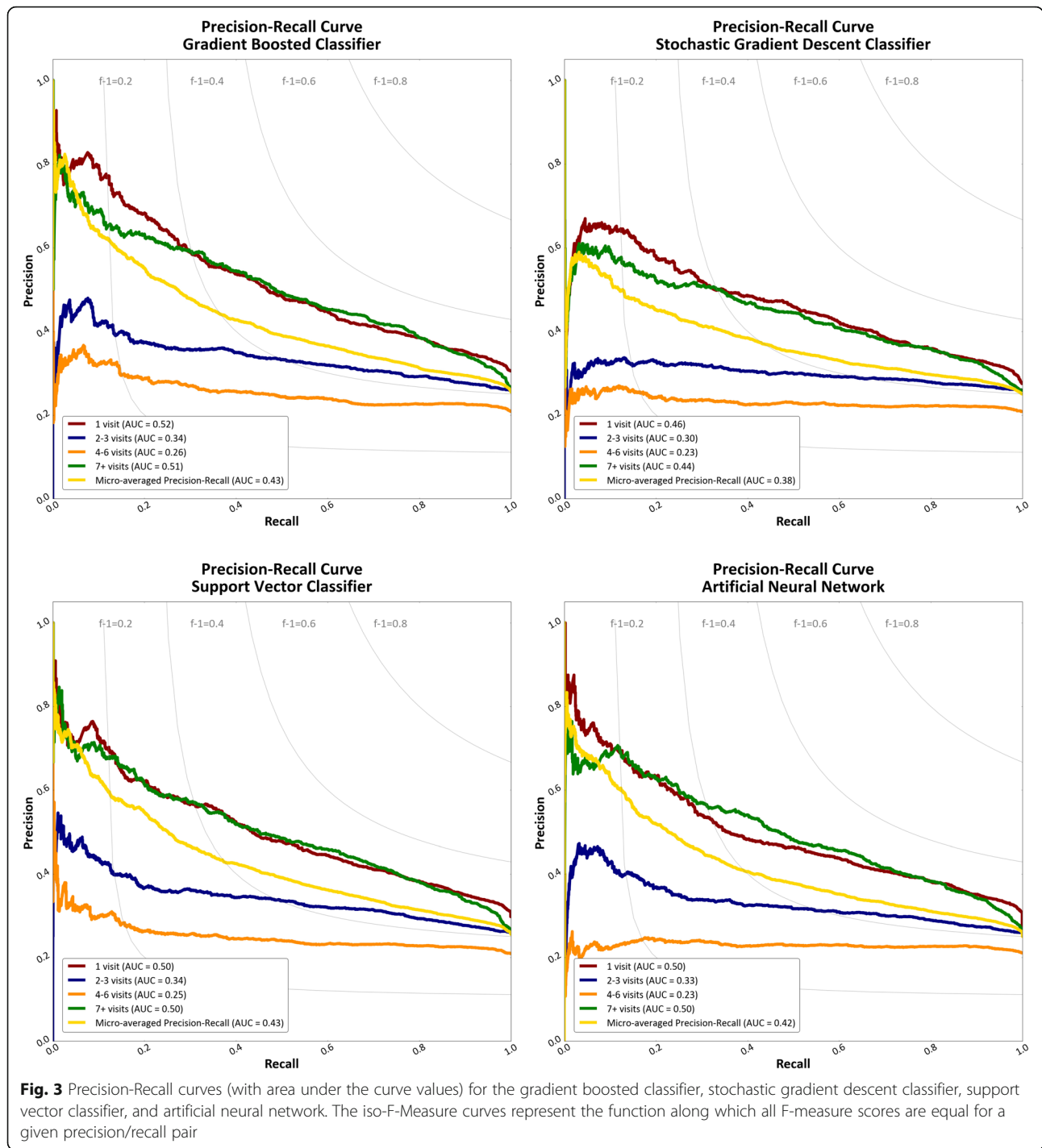
half of those that are truly in a class are correctly classified as such, resulting in many false negatives.

Difficulty in accurately predicting one-year chiropractic service utilization could be expected based on the existing body of literature and the correlation analysis performed during this study. Machine learning using complex pain-related data may help to identify phenotypic subsets of pain presentations and uncover previously unidentifiable relationships between important variables, including factors related to pain-related healthcare service utilization [40]. However, clinical or sociodemographic features correlating with higher or lower service utilization have yet to be demonstrated quantitatively and empirically. In support of this, we found no features demonstrating a strong correlation to one-year chiropractic service utilization. It is likely that the features included in this dataset may be poor predictors of service utilization and more potentially relevant data may exist, such as facility or clinic characteristics, to yield more accurate predictive capabilities. For instance, a short supply of chiropractic appointments available may impose ceiling limits on the number of visits available to a given patient, irrespective of any optimal amount. If availability were uniform across all sites and all time points, then it is possible the features included may have different predictive abilities. We find that while there may be limited explanatory value in our included variables with respect to pattern recognition of one-year chiropractic service utilization, there may be additional explanatory value of these variables with respect to recognition of other clinically meaningful patterns which warrant future investigation.

Although limited evidence currently exists to suggest a relationship between optimal amount of chiropractic care and clinical outcomes, the clinical implications of label misclassification may result in over- or underestimation of service utilization, with both yielding potentially increased front-end or back-end system burden. Overestimating service utilization may result in decreased access to clinically indicated care for other veterans due to an overburdening of the system on the front-end through inappropriate resource allocation. Further, high rates of service overuse may substantially contribute to higher healthcare spending and may result in harms to all stakeholders in the healthcare system, especially patients [41]. Underestimating service utilization may result in an individual veteran failing to receive an adequate amount of chiropractic care that may be clinically indicated. This may cause an increase in back-end system burden in that the individual veteran may require additional resource allocation than previously predicted.

One strength of this study is that it uses one of the largest cohorts of patients receiving chiropractic care within a capitated healthcare system. We also examine the use of chiropractic services on a rolling basis over a



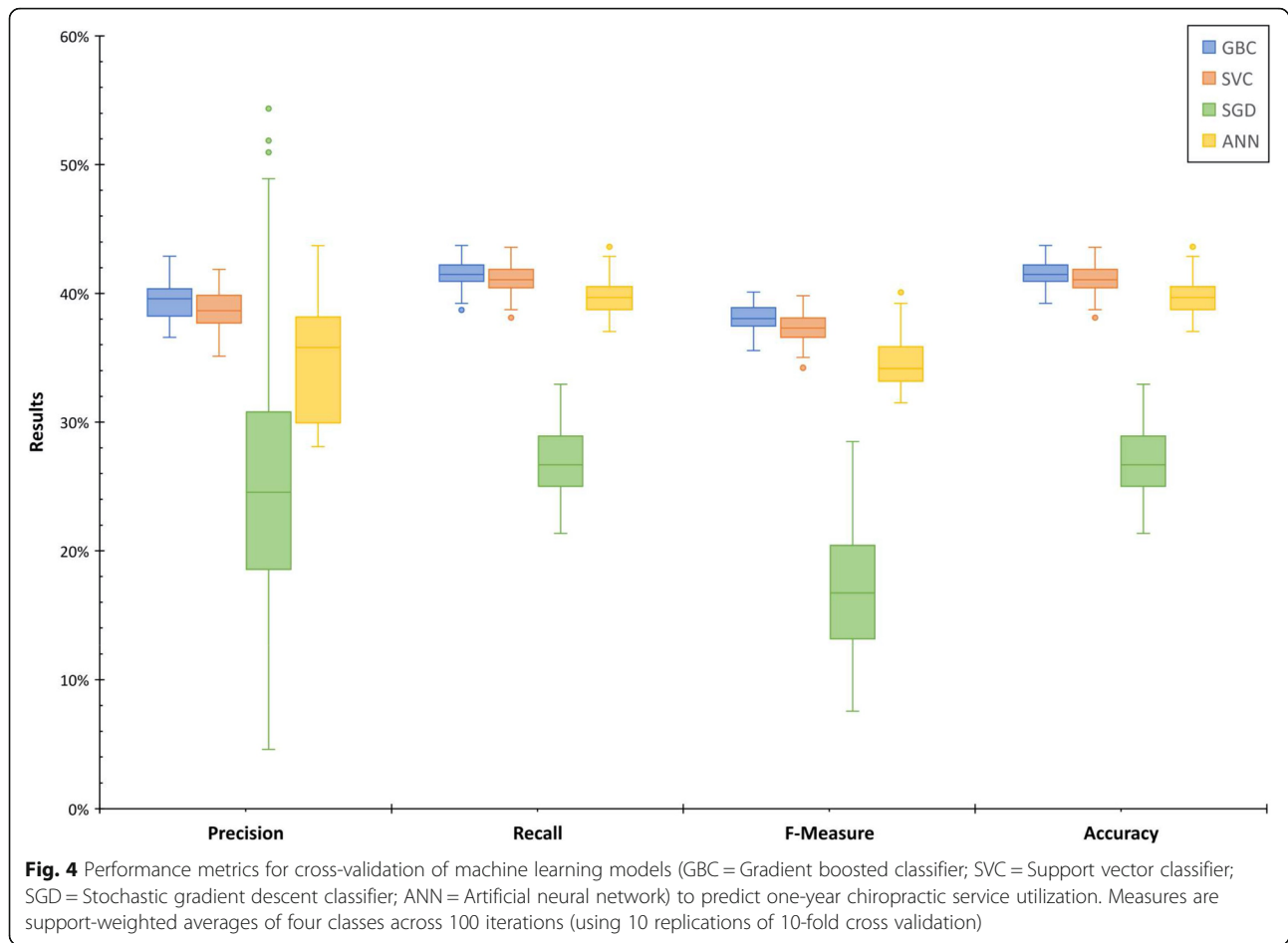


twelve-year study period, allowing for analyses that reflect the development of VA on-station chiropractic clinics over time.

One of the most interesting findings of this study is the distribution of the chiropractic service utilization into quartiles. While empirical evidence regarding optimal treatment trial duration is limited, the recommended chiropractic treatment frequency and duration

for VA patients with spine-related symptoms is up to 10 visits for uncomplicated acute episodes and up to 12 visits for complicated acute episodes and chronic conditions, based on Delphi consensus processes [26]. These data suggest that VA chiropractors have been providing care consistent with these recommendations.

For the 25% of veterans receiving 7 or more chiropractic care visits within 1 year, it is unclear if these visits



were related to a single episode of care for a condition or related to a multiple episodes of care for either a new condition or reactivation of an existing condition throughout the one-year period. This quartile included the widest spread of visits (7 to 73 visits), with less than 1% of all patients receiving more than 24 visits. It is possible that patients with greater service use are of higher medical complexity with higher rates of medical and mental health comorbidities. This is consistent with previous work demonstrating higher service utilization associated with higher comorbidity burdens in the veteran population [17]. These features of complexity may be represented in our dataset, thus contributing to our models for slightly improved pattern recognition in the class of highest service use. Further, there may be a time-dependent underlying relationship between service utilization and specific facilities related to their individual clinic characteristics. Some facilities may have greater capacity for clinic visits based on clinic characteristics such as physical space and number of chiropractors, which could have a greater influence on chiropractic service utilization than patient characteristics. Additionally, provider-based factors beyond the scope of this project – such as the

influence of job performance metrics and/or chiropractic practice preferences – may impact the number of visits that patients receive.

Additionally, 25% of veterans received only a single visit of VA chiropractic services. This may be confounded due to inappropriateness of chiropractic care for the veteran's presenting condition, veteran's preference to not seek additional chiropractic care, or a single on-site consultation followed by referral for purchased care off-site [10]. The clinical determination of inappropriateness of chiropractic care and the relationship between supply of chiropractic services by facility may be related to features included in our dataset, thus supporting pattern recognition for this class of lowest service use.

There are several limitations to this study. First, this study was performed to examine VA chiropractic service utilization specifically, which may limit generalizability to chiropractic clinics outside of VA. We used clinical data from the fully-integrated VA EHR that may not be easily obtained in private practice chiropractic clinics.

We relied on administrative data from the VA EHR as an abstraction of clinical data. We did not seek to examine what occurred at individual chiropractic visits, which

may have impacted our findings. More detailed analyses of visits, including natural language processing of progress note documentation, critical evaluation of treatments provided, and inclusion of data from valid and reliable patient-reported outcome measures may provide more detailed clinical data and improve future predictive models.

We did not examine any differences in service utilization based on specific pain presentation because a substantial majority of patients in our dataset (86%) received care for low back pain alone or in combination with neck pain. We included concurrent neck pain and concurrent other musculoskeletal pain at both the index chiropractic visit and across the one-year observation period as features in our model to account for potential mediation of these on service utilization. We hypothesize little change in our classification performance based on these factors alone. Currently, there is limited literature available regarding the optimal frequency and duration of chiropractic care recommended for these specific pain presentations [24], making it unlikely that there are system-wide patterns in chiropractic service utilization based on pain presentation.

The sociodemographic and clinical data used for each veteran was based on those collected at the time of his or her MSD Cohort entry, with comorbidity diagnoses occurring from 12 months prior to 6 months after cohort entry. By limiting inclusion criteria to veterans with an index chiropractic date within 365 days of cohort entry, we attempted to limit potential inaccuracy of these data. However, it remains possible that a veteran's health status may have changed over the 365 days following cohort entry and/or the 365 days following the index chiropractic visit. It is also possible a veteran may have presented for his or her index chiropractic visit within 6 months after cohort entry and prior to a diagnosis of a comorbidity.

We aimed to predict visit quartile as a multiclass classification problem, with our results suggesting limited clinical utility to this approach. Different results may be found by structuring the question as a binary classification problem (for example, classifying patients based on a certain clinically relevant threshold of visits) or as a regression problem (predicting service utilization across a continuous quantity of visits). We also used a sampling of commonly used classification models to predict one-year chiropractic service utilization, with a trial-and-error grid-search approach to hyperparameter tuning. It is possible, although we suspect minimally likely, that other functions and/or hyperparameters may yield stronger classification performance using these same data.

We selected a 70–30% training-testing split for our initial model development. It is possible that different results may be found by training on a larger proportion of

the dataset (i.e. 80% or 90%) with a smaller testing set. However, given the similar results of our 10-fold cross validation (with a 10% testing set for each fold), increasing the size of our training set is unlikely to strongly change our classification performance.

Specific to the algorithms selected, we identified a warning in the executed Python code that was present in both the stochastic gradient descent classifier and the artificial neural network. During the cross-validation phase, both algorithms resulted in zero instances of a predicted class during a small number of iterations. We recognize this as a limitation in spite of using 10-fold repeated, stratified cross validation, with it possible that an individual class may not be predicted in some iterations of these models. Additional calibration of these models to better predict labels more consistent with baseline probabilities may help to address this. This may have contributed to the weaker and less consistent performance of these two models compared to the gradient boosted classifier and support vector classifier.

## Conclusion

Overall, we have demonstrated that using supervised machine learning to predict chiropractic service utilization remains challenging. Preliminary performance shows a small shift in predictive probability over naïve classification. However, model performance metrics suggest limited clinical utility at this time based on the features included in our dataset. Future work should examine mechanisms to improve model performance, including collecting potentially more relevant data such as facility and clinic access characteristics, progress note documentation, treatments rendered, and patient-reported outcome measures.

## Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12998-020-00335-4>.

**Additional file 1.** Variables included in the final dataset.

**Additional file 2.** Detailed description of methods, including machine learning algorithms used.

## Acknowledgements

The contents of this manuscript represent the view of the authors and do not necessarily reflect the position or policy of the U.S. Department of Veterans Affairs or the United States Government.

## Authors' contributions

BCC, SF, AJL, JLG, KLC, and CAB were responsible for the design and conception of this study. BCC, SF, JLG, KLC, HB, and CAB were responsible for data acquisition and analysis. All authors were responsible for interpretation of results and critical revision of the manuscript. All authors read and approved of the final manuscript.

## Funding

This material is based upon work supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Academic Affiliations, Office of Research and Development, and Health Services Research and

Development IIR-16-262 (Goulet, PI), IIR-12-118, and CIN-13-407, with resources and the use of facilities at the VA Connecticut Healthcare System, West Haven, CT. The authors have no additional conflicts of interest, financial or otherwise, to disclose.

#### Availability of data and materials

To maximize protection security of veterans' data while making these data available to researchers, the US Department of Veterans Affairs (VA) developed the VA Informatics and Computing Infrastructure (VINCI). VA researchers must log onto VINCI via a secure gateway or virtual private network connection (VPN), and use a virtual workspace on VINCI to access and analyze VA data. By VA Office of Research and Development policy, VINCI does not allow the transfer of any patient-level data out of its secure environment without special permission. Researchers who are not VA employees must be vetted and receive "without compensation" (WOC) employee status to gain access to VINCI. All analyses performed for this study took place on the VINCI platform. For questions about data access, contact the study lead ([Brian.Coleman2@va.gov](mailto:Brian.Coleman2@va.gov)) or the VA Office of Research and Development ([VHACOORDRegulatory@va.gov](mailto:VHACOORDRegulatory@va.gov)).

#### Ethics approval and consent to participate

This study was approved by the Institutional Review Board of the VA Connecticut Healthcare System, under continuing review of the Musculoskeletal Diagnosis Cohort Study (#0005, PI: Goulet, Brandt).

#### Consent for publication

Not Applicable.

#### Competing interests

The authors declare that they have no competing interests.

Received: 18 February 2020 Accepted: 2 July 2020

Published online: 17 July 2020

#### References

- Gaskin DJ, Richard P. The economic costs of pain in the United States. *J Pain*. 2012;13:715–24.
- Dahlhamer J, Lucas J, Zelaya C, Nahin R, Mackey S, DeBar L, Kerns R, Von Korff M, Porter L, Helmick C. Prevalence of chronic pain and high-impact chronic pain among adults - United States, 2016. *MMWR Morb Mortal Wkly Rep*. 2018;67:1001–6.
- Herman PM, Broten N, Lavelle TA, Sorbero ME, Coulter ID. Health care costs and opioid use associated with high-impact chronic spinal pain in the United States. *Spine (Phila Pa 1976)*. 2019;44:1154–61.
- Hartvigsen J, Hancock MJ, Kongsted A, Louw Q, Ferreira ML, Genevay S, Hoy D, Karpinen J, Pransky G, Sieper J, Smeets RJ, Underwood M and Lancet Low Back Pain Series Working Group. What low back pain is and why we need to pay attention. *Lancet*. 2018;391:2356–67.
- Herman PM, Yuan AH, Cefalu MS, Chu K, Zeng Q, Marshall N, Lorenz KA, Taylor SL. The use of complementary and integrative health approaches for chronic musculoskeletal pain in younger US veterans: an economic evaluation. *PLoS One*. 2019;14:e0217831.
- Herman PM, Poindexter BL, Witt CM, Eisenberg DM. Are complementary therapies and integrative care cost-effective? A systematic review of economic evaluations. *BMJ Open*. 2012;2:e001046.
- Dagenais S, Brady O, Haldeman S, Manga P. A systematic review comparing the costs of chiropractic care to other interventions for spine pain in the United States. *BMC Health Serv Res*. 2015;15:474.
- Liu X, Hanney WJ, Masaracchio M, Kolber MJ, Zhao M, Spaulding AC, Gabriel MH. Immediate physical therapy initiation in patients with acute low back pain is associated with a reduction in downstream health care utilization and costs. *Phys Ther*. 2018;98:336–47.
- Herman PM, Lavelle TA, Sorbero ME, Hurwitz EL, Coulter ID. Are nonpharmacologic interventions for chronic low back pain more cost effective than usual care? Proof of concept results from a Markov Model. *Spine (Phila Pa 1976)*. 2019;44:1456–64.
- Lisi AJ, Brandt CA. Trends in the use and characteristics of chiropractic Services in the Department of Veterans Affairs. *J Manip Physiol Ther*. 2016; 39:381–6.
- Becker WC, DeBar LL, Heapy AA, Higgins D, Krein SL, Lisi A, Makris UE, Allen KD. A research agenda for advancing non-pharmacological management of chronic musculoskeletal pain: findings from a VHA state-of-the-art conference. *J Gen Intern Med*. 2018;33:11–5.
- Qaseem A, Wilt TJ, McLean RM, Forcica MA and Clinical Guidelines Committee of the American College of Physicians. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. *Ann Intern Med*. 2017;166:514–30.
- Kligler B, Bair MJ, Banerjee R, DeBar L, Ezeji-Okoye S, Lisi A, Murphy JL, Sandbrink F, Cherkin DC. Clinical policy recommendations from the VHA state-of-the-art conference on non-pharmacological approaches to chronic musculoskeletal pain. *J Gen Intern Med*. 2018;33:16–23.
- Goulet JL, Kerns RD, Bair M, Becker WC, Brennan P, Burgess DJ, Carroll CM, Dobscha S, Driscoll MA, Fenton BT, Fraenkel L, Haskell SG, Heapy AA, Higgins DM, Hoff RA, Hwang U, Justice AC, Piette JD, Sinnott P, Wandner L, Womack JA, Brandt CA. The musculoskeletal diagnosis cohort: examining pain and pain care among veterans. *Pain*. 2016;157:1696–703.
- Higgins DM, Kerns RD, Brandt CA, Haskell SG, Bathulapalli H, Gilliam W, Goulet JL. Persistent pain and comorbidity among Operation Enduring Freedom/Operation Iraqi Freedom/Operation New Dawn veterans. *Pain Med*. 2014;15:782–90.
- Coleman BC, Corcoran KL, DeRycke EC, Bastian LA, Brandt CA, Haskell SG, Heapy AA, Lisi AJ. Factors associated with posttraumatic stress disorder among veterans of recent wars receiving Veterans Affairs chiropractic care. *J Manipulative Physiol Ther*. 2020;S0161-4754(20):30064–6.
- Beehler GP, Rodrigues AE, Mercurio-Riley D, Dunn AS. Primary care utilization among veterans with chronic musculoskeletal pain: a retrospective chart review. *Pain Med*. 2013;14:1021–31.
- Simpao AF, Ahumada LM, Galvez JA, Rehman MA. A review of analytics and clinical informatics in health care. *J Med Syst*. 2014;38:45.
- Janke AT, Overbeek DL, Kocher KE, Levy PD. Exploring the potential of predictive analytics and big data in emergency care. *Ann Emerg Med*. 2016; 67:227–36.
- Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med*. 2016;23:269–78.
- Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *J Biomed Inform*. 2014;48:160–70.
- Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, Delitto A. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. *Ann Intern Med*. 2004;141:920–8.
- Dougherty PE, Karuza J, Savino D, Katz P. Evaluation of a modified clinical prediction rule for use with spinal manipulative therapy in patients with chronic low back pain: a randomized clinical trial. *Chiropr Man Therap*. 2014;22:41.
- Pasquier M, Daneau C, Marchand AA, Lardon A, Descarreaux M. Spinal manipulation frequency and dosage effects on clinical and physiological outcomes: a scoping review. *Chiropr Man Therap*. 2019;27:23.
- Globe G, Farabaugh RJ, Hawk C, Morris CE, Baker G, Whalen WM, Walters S, Kaeser M, Dehen M, Augat T. Clinical practice guideline: chiropractic care for low back pain. *J Manip Physiol Ther*. 2016;39:1–22.
- Lisi AJ, Salisbury SA, Hawk C, Vining RD, Wallace RB, Branson R, Long CR, Burgo-Black AL, Goertz CM. Chiropractic integrated care pathway for low back pain in veterans: results of a Delphi consensus process. *J Manip Physiol Ther*. 2018;41:137–48.
- Justice AC, Erdos J, Brandt C, Conigliaro J, Tierney W, Bryant K. The Veterans Affairs Healthcare System: a unique laboratory for observational and interventional research. *Med Care*. 2006;44:57–12.
- Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323.
- Sinnott PL, Siroka AM, Shane AC, Trafton JA, Wagner TH. Identifying neck and back pain in administrative data: defining the right cohort. *Spine (Phila Pa 1976)*. 2012;37:860–74.
- Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. *Anesth Analg*. 2018;126:1763–8.
- Saeyns Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;19:2507–17.
- Lever J, Kryzwiniski M, Altman N. Principal component analysis. *Nat Methods*. 2017;14:641–2.

33. Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. *Biol Direct*. 2007;2:2.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
35. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10:e0118432.
36. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. *Ann Fam Med*. 2014;12:573–6.
37. Hu Z, Hao S, Jin B, Shin AY, Zhu C, Huang M, Wang Y, Zheng L, Dai D, Culver DS, Alfreds ST, Rogow T, Stearns F, Sylvester KG, Widen E, Ling X. Online prediction of health care utilization in the next six months based on electronic health record information: a cohort and validation study. *J Med Internet Res*. 2015;17:e219.
38. Jones A, Costa AP, Pesevski A, McNicholas PD. Predicting hospital and emergency department utilization among community-dwelling older adults: statistical and machine learning approaches. *PLoS One*. 2018;13:e0206662.
39. Rosella LC, Kornas K, Yao Z, Manuel DG, Bornbaum C, Fransoo R, Stukel T. Predicting high health care resource utilization in a single-payer public health care system: development and validation of the high resource user population risk tool. *Med Care*. 2018;56:e61–9.
40. Lotsch J, Ullsch A. Machine learning in pain research. *Pain*. 2018;159:623–30.
41. Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, Nagpal S, Saini V, Srivastava D, Chalmers K, Korenstein D. Evidence for overuse of medical services around the world. *Lancet*. 2017;390:156–68.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

