



ScienceDirect

Contents lists available at [sciencedirect.com](http://sciencedirect.com)  
Journal homepage: [www.elsevier.com/locate/jval](http://www.elsevier.com/locate/jval)

## ABSTRACT

### Advanced Analytics Studies

#### A11

##### A MACHINE LEARNING APPROACH TO PREDICTING MORTALITY IN CYSTIC FIBROSIS

Rodríguez PJ, Heagerty PJ, Goss CH, Veenstra DL, Bansal A  
*University of Washington, Seattle, WA, USA*

**Objectives:** Cystic Fibrosis (CF) is a progressive genetic disease affecting the lungs. Referral for lung transplant (LTx) is recommended when patients face a high likelihood of short-term mortality. Current models for predicting mortality in CF have poor performance, making LTx referral decision-making difficult and inconsistent in practice. Our objective is to apply machine learning (ML) methods to develop a high-performing risk prediction model for short-term mortality in CF. **Methods:** We used data from the Cystic Fibrosis Foundation Patient Registry to develop a mortality risk prediction model for adults with CF. We used the lasso method to develop a preliminary ML model on a limited initial set of clinical and demographic variables. We evaluated the time-varying performance of baseline predictions using the area under the receiver operating curve (AUC) over time and summarized 2-year performance using the survival concordance index (c-index). Performance was compared to the existing model, forced expiratory volume in 1 second (FEV<sub>1</sub>) alone. **Results:** Our lasso model identified 24 predictors of mortality from the initial limited dataset and had a higher AUC at all time points compared to the existing FEV<sub>1</sub> only model (c-index 0.89(0.86 - 0.92) vs 0.85(0.81 - 0.88)). **Conclusions:** The lasso predicted mortality better than FEV<sub>1</sub> alone for adults with CF in the US using a preliminary dataset. We are now training models on an expanded dataset using additional ML approaches, including ridge, elastic net, support vector machines, random forests, and boosting. Instead of choosing only one ML model, we will create an optimally weighted combination of these different models using ensemble learning methods. We hypothesize that even greater gains in performance will be achieved.



#### A12

##### PREDICTORS OF PARKINSON DISEASE IN A MEDICARE POPULATION: AN APPLICATION OF MACHINE LEARNING IN EARLY DISEASE DETECTION

Icten Z, Burns SM, Menzin JA  
*Boston Health Economics, Boston, MA, USA*

**Objectives:** Detecting Parkinson disease (PD) early in its prodromal period can facilitate timely treatment and mitigate symptoms and risks. This study aims to identify predictors of incident PD patients using Medicare administrative claims data and machine learning techniques. **Methods:** Medicare Part A/B claims (5% sample) from 2010Q1-2015Q4 were used to identify incident PD cases in 2015 based on ICD-9/10 diagnosis codes in patients  $\geq 65$  years old with continuous enrollment during the two-year baseline period prior to their PD diagnosis date (index date). Controls were identified in a 3:1 ratio to cases and met the same eligibility criteria, had no evidence of PD diagnosis and had a randomly selected index encounter date with the month/year matching to the case's index date. Features included demographics, comorbidities, medication and procedure utilization, and service location variables were extracted from the baseline. Data were partitioned using a 60%/20%/20% split to train, tune models and test performance on unseen data. Traditional and regularized logistic regression, k-nearest neighbor, XGBoost, support vector machine, and random forest models were built, and the best model was selected using the area under the ROC curve (AUC). Accuracy, recall, precision and F1 score were also assessed. **Results:** The study population included 4,575 cases and 13,725 controls (mean age=74.8 years; females=56%). The XGBoost model was the best performing model (on unseen data: AUC: 83.1%; accuracy: 79.6%; recall: 65.1%; precision: 58.2% and F1: 0.61). Features in the XGBoost model with the highest weights were age, gender, Charlson Comorbidity Index, motor symptoms (tremor, abnormal gait, involuntary movements), diagnoses related to autonomic dysfunction (urinary incontinence), diagnostic procedures (CT scan, MRI), hypertension, cancer diagnostics (mammography), ophthalmologic disease, anxiety and psychosis. **Conclusions:** Our study identified



predictors of PD with high predictive accuracy and our findings are overall consistent with established risk factors, while indicating opportunities for further research.

#### A13

##### PREDICTION OF BREAST CANCER USING K-NEAREST NEIGHBOUR: A SUPERVISED MACHINE LEARNING ALGORITHM

Pandey S,<sup>1</sup> Sharma A,<sup>1</sup> Siddiqui MK,<sup>1</sup> Singla D,<sup>2</sup> Vanderpuye-Orgle J<sup>3</sup>  
<sup>1</sup>Parexel International, Mohali, PB, India, <sup>2</sup>Parexel International, Bangalore, India, <sup>3</sup>PAREXEL, Glendale, CA, USA

**Objectives:** Mammograms are not 100% accurate in the identifying the breast cancer. Better methods are needed to predict the breast cancer without the need of surgical biopsies. The study evaluated the prediction accuracy of breast cancer using the K-nearest neighbor (k-NN) classifier algorithm. **Methods:** The breast cancer dataset (containing 569 records and 32 attributes) was obtained from University of California Irvine (UCI) machine learning repository. Applying supervised machine learning technique to patient characteristics including tumor features (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension), k-NEAREST NEIGHBOUR (k-NN) was used to detect whether the mass was malignant or benign. Data were segmented into a training dataset containing the first 469 observations to build the k-NN model and a testing dataset containing the remaining observations was used to simulate new patients. Normalization of the data points was applied to rescale the features to a standard range of values. The initial choice of  $k = 21$ , approximately square root of 469 patients in our training dataset was used. Alternative k-values ( $k = 1, 5, 11, 15, 21, 27$ ) were also tested to optimize the model performance. The analysis was conducted using "class" package of R (v3.6.2). **Results:** In 100 simulations, 98% accuracy was achieved by the k-NN algorithm – i.e., only 2 out of 100, or 2 percent of masses were incorrectly classified. Choice of  $k=21$  seems more accurate than any other choices as it has the minimum number of incorrect identification of cancerous cells. **Conclusions:** Supervised machine learning algorithm was shown to be capable of tackling extremely complex tasks such as identification of cancerous masses with reasonable accuracy. The application of this analysis could be an important resource for early detection of cancerous tumors and their treatment.



#### A14

##### CAN WE DECREASE THE SCREENING BURDEN IN SYSTEMATIC REVIEWS? PERFORMANCE OF TWO NATURAL LANGUAGE PROCESSORS TO EXCLUDE RECORDS

Tetzlaff J,<sup>1</sup> Murad MH,<sup>2</sup> Wang Z<sup>2</sup>  
<sup>1</sup>Evidence Partners, Kanata, ON, Canada, <sup>2</sup>Mayo Clinic Evidence-Based Practice Center, Rochester, MN, USA

**Objectives:** Systematic reviews (SR) require both comprehensive and efficient methodology and automation helps address these competing conditions. Natural language processors (NLP) may decrease screening time (e.g. prioritized screening); however, their benefits/risks as a second or autonomous reviewer require more investigation. This study assessed the performance of NLPs to exclude records from SRs and compared their performance to human reviewers. **Methods:** Using data from 8 completed SRs conducted by a US Evidence-Based Practice Center, we randomly selected 10% of references from each SR and trained two NLP classifiers (one support vector machine (SVM) and one Naïve Bayes (NB)) on inclusion/exclusion decisions. The classifiers screened the remaining references, leaving records unreviewed in the absence of consensus. Our primary outcome was NLP false negative (FN) rate (FN/screened records) compared to dual human-screened results. We also estimated single-human FN rates, workload savings and the potential impact of NLP FN on review results/conclusions. **Results:** Including 33,191 total screened records, the SRs spanned diverse topic domains (e.g. metabolic, neoplasms, respiratory), review types (e.g. interventional, umbrella, qualitative) and proportions of included studies. NLP FN rates ranged from 0 to 0.04%;

