International Journal of Population Data Science

Journal Website: www.ijpds.org

# Machine learning for identification of frailty in Canadian primary care practices

Sylvia Aponte-Hao[1,*], Sabrina T. Wong[2], Manpreet Thandi[2], Paul Ronksley[1], Kerry McBrien[1], Joon Lee[1], Mathew Grandy[3], Alan Katz[4], Dee Mangin[5], Alexander Singer[6], Donna Manca[7], and Tyler Williamson[1]

**Submission History**

| | |
|---|---|
| Submitted: | 18/12/2021 |
| Accepted: | 02/07/2021 |
| Published: | 09/09/2021 |

[1]University of Calgary, Cumming School of Medicine, 3330 Hospital Drive NW, Calgary, Alberta, T2N 4N1
[2]University of British Columbia, Centre for Health Services and Policy Research & School of Nursing, 2211 Wesbrook Mall, Vancouver, BC, V6T 2B5
[3]Department of Family Medicine, Dalhousie University, 1465 Brenton Street, Suite 402, Halifax, Nova Scotia, B3J 3T4
[4]College of Medicine Faculty of Health Sciences, University of Manitoba, 408-727 McDermot Ave, Winnipeg, Mb, R3E 3P5
[5]Department of Family Medicine, McMaster University, 1280 Main St W, Hamilton, ON, L8S 4L8
[6]Department of Family Medicine, University of Manitoba, 408-727 McDermot Ave, Winnipeg, Mb, R3E 3P5
[7]Department of Family Medicine, University of Alberta, 610 University Terrace, 8303 - 112 Street NW, Edmonton, Alberta, T6G 2T4

## Abstract

### Introduction
Frailty is a medical syndrome, commonly affecting people aged 65 years and over and is characterized by a greater risk of adverse outcomes following illness or injury. Electronic medical records contain a large amount of longitudinal data that can be used for primary care research. Machine learning can fully utilize this wide breadth of data for the detection of diseases and syndromes. The creation of a frailty case definition using machine learning may facilitate early intervention, inform advanced screening tests, and allow for surveillance.

### Objectives
The objective of this study was to develop a validated case definition of frailty for the primary care context, using machine learning.

### Methods
Physicians participating in the Canadian Primary Care Sentinel Surveillance Network across Canada were asked to retrospectively identify the level of frailty present in a sample of their own patients (total n = 5,466), collected from 2015–2019. Frailty levels were dichotomized using a cut-off of 5. Extracted features included previously prescribed medications, billing codes, and other routinely collected primary care data. We used eight supervised machine learning algorithms, with performance assessed using a hold-out test set. A balanced training dataset was also created by oversampling. Sensitivity analyses considered two alternative dichotomization cut-offs. Model performance was evaluated using area under the receiver-operating characteristic curve, F1, accuracy, sensitivity, specificity, negative predictive value and positive predictive value.

### Results
The prevalence of frailty within our sample was 18.4%. Of the eight models developed to identify frail patients, an XGBoost model achieved the highest sensitivity (78.14%) and specificity (74.41%). The balanced training dataset did not improve classification performance. Sensitivity analyses did not show improved performance for cut-offs other than 5.

### Conclusion
Supervised machine learning was able to create well performing classification models for frailty. Future research is needed to assess frailty inter-rater reliability, and link multiple data sources for frailty identification.

### Keywords
electronic medical records; electronic health records; machine learning; supervised machine learning; case definition; frailty; primary care; Canada

*Corresponding Author:
*Email Address:* zhi.hao@ucalgary.ca (Sylvia Aponte-Hao)

https://doi.org/10.23889/ijpds.v6i1.1650

# Introduction

Frailty is a medical syndrome, commonly affecting people aged 65 years and older, characterized by a greater risk of adverse outcomes following illness or injury, despite accounting for age, other diseases, and medical treatment [1]. Frailty is associated with higher health care costs [2], greater risk of adverse events during [3] and post-surgery [4], markedly worse quality of life [5] and increased burden for family caregivers of frail patients [6]. As of 2018, there were an estimated 1.5 million Canadians living with frailty [7], and by 2025 this number is projected to increase to over 2 million. However, studies have demonstrated that frailty can be delayed or improved through a variety of interventions, such as nutrient supplementation and increased exercise [8]. Primary care is often the first point of care for patients and thus accurate identification of frailty in this setting may enable improved management of identified individuals such as ensuring early initiation of interventions [9] and informing advanced frailty screening tests [10], which could lead to reduced downstream costs through reduced hospitalizations [9]. There is currently no standard definition or instrument to measure frailty, and frailty prevalence estimates have found to vary greatly depending on the frailty instrument used [11].

Electronic medical records (EMRs) are a rich clinical data source for primary care research. Disease case definitions are routinely created and validated for the identification of patient cohorts. Machine learning has successfully been used in the creation of case definitions for other diseases such as hypertension and osteoarthritis in primary care that are being used for practice reporting, quality improvement, public health surveillance, and research [12]. Previous work using supervised machine learning for the identification of frailty in EMR data by Williamson et al. used data from Alberta, Canada [13]. This study defined frailty using the Clinical Frailty Scale [14] showed fair performance, achieving a sensitivity of 0.28 and a specificity of 0.94. Other research has been done on the classification of frailty using machine learning methods, but frailty was defined using other instruments. Hassler et al. identified frailty using the Frailty Phenotype, while also using supervised machine learning methods but not using EMR data [15]. This research obtained sensitivity estimates ranging between 65.7% to 86.7%, and specificity ranging between 58.1% to 85.6% [16]. Ambagtsheer et al. used the electronic Frailty Index [17] for the identification of frailty, while also using supervised machine learning methods on EMR data [18]. The best performing model was able to achieve a sensitivity of 97.8% and a specificity of 89.1%.

The objective of this study was to develop a validated case definition of frailty for the primary care context using machine learning. The creation of a frailty case definition using supervised machine learning for wide distribution and deployment in Canadian primary care practices may allow for surveillance of frailty, future research on frail cohorts, as well as contribute to better management and care for frail patients.

# Methods

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a pan-Canadian primary care database that routinely collects and stores de-identified patient EMR data from eight provinces and one territory across Canada, with approximately 1.8 million unique patients in its database, and over a thousand primary care providers [19]. Data consistently captured within CPCSSN include diagnoses, billing codes, and prescribed medications.

## Reference set creation

Five CPCSSN networks participated in data collection and building the reference set of frail patients: British Columbia, Alberta, Nova Scotia, Ontario, and Manitoba. Primary care physicians used the Rockwood Clinical Frailty Scale (CFS) to retrospectively classify the degree of frailty observed in their patients aged 65 and above who were seen within the last 24 months. The CFS is a validated frailty measurement tool commonly used in primary care and is based on short written descriptions of increasing levels of frailty [14] (Appendix I). The CFS ranges from 1 to 9, with 1 having the label of 'very fit' and 9 labelled 'terminally ill' (the highest degree of frailty).

Physicians were given the option of basing their assessments on recall or querying information in their EMRs in order to provide an accurate frailty rating. Data were gathered in two stages, with the initial data collection being restricted to Alberta only. Data from Alberta were gathered in 2015 for a previous study that focused on frailty identification [13], while the other provincial sites gathered data in 2019. A total of 5,466 patients were rated by 90 physicians in total across the five regional CPCSSN sites located across five Canadian provinces, with each patient receiving one CFS rating by their physician only.

## Feature engineering

We extracted patient EMR data from participating CPCSSN sentinels with the accompanying CFS score and included features (measured data elements representative of a patient characteristic) that were present in all EMRs to form a unified dataset. Patient visit diagnoses, prescribed medications, lab results and biometrics (height, weight and body mass index) were extracted from the past two years prior to the CFS score assignment. For the purposes of this study, frailty was dichotomized into frail or not frail from the original physician-rated CFS score, with those receiving a score of 5 or higher on the CFS being labelled as frail, and those with a score of 4 or lower being labelled as not frail.

The following chronic conditions, detected using validated case-detection algorithms available in CPCSSN [20], were extracted: chronic obstructive pulmonary disease, dementia, depression, epilepsy, hypertension, and osteoarthritis. Patient demographics such as age and sex were also extracted.

We performed feature selection by removing features with very low variability, as defined by the ratio of the most common value to the second most common value being more than a ratio of 95:5. As missing data are often observed in EMR data, any feature with greater than 20% missing data was also removed (with the exception of height, weight, and BMI information as these were considered to be potentially important), and for those with less than 20% missing data (systolic and diastolic blood pressures), single imputation using predictive mean matching [21] was performed. In addition

to single imputation, missingness-indicator variables were also created, in the event that these missing values were potentially related to frailty status. The regional CPCSSN network the data were collected from was also included as a feature, as there may be regional data extraction and processing differences. However, the inclusion of this feature limits the generalizability of the models to only the regional CPCSSN networks included in this study.

After the removal of features with low variance or high correlation, no additional feature selection was performed as this had reduced the total number of features from 5,466 to 75. The final set of features used is presented in Table 1.

## Supervised machine learning

Patients were partitioned into a 30-70 split, with 70% (n = 3,827) of patients used for training, and 30% (n = 1,639) as the hold-out test set. Within the training set, there were 3,103 non-frail patients and 724 frail patients. The hold-out test set had 1,360 non-frail patients and 279 frail patients.

Numeric features in both the training and test sets were scaled and centered according to the training set to ensure no data leakage from the validation set. Within the 70% training set, the data were split into five random folds for cross validation to guard against overfitting.

Imbalanced data can result in biased estimates of training performance, especially when the class of interest is the minority class. A model predicting everyone as non-frail will still result in a 81.1% accuracy rate, but is actually of no value when none of the frail patients have been identified. One method of combating imbalanced datasets is to oversample the minority class such that the training data becomes balanced. As the original training data were imbalanced (18.9% frail), synthetic minority over-sampling technique (SMOTE) [22] was also performed to create synthetic samples of frail patients, such that there were an equal number of frail (n = 3,103) and non-frail (n = 3,103) patients. The specific implementation of SMOTE used was SMOTE-Nominal Continuous (SMOTE-NC), as there are both categorical and numerical data as features (for the sake of simplicity, future references of SMOTE-NC will be simplified to just SMOTE). SMOTE was used only to create more synthetic frail observations, and no undersampling of non-frail patients was performed.

A random search of 60 combinations was used for hyperparameter tuning within five fold cross validation, with the best performing model chosen by average sensitivity across the five folds for the balanced training dataset created by SMOTE, and average area under the receiver-operating characteristic curve (AUC) for the imbalanced datasets.

A selection of seven commonly used binary supervised machine learning architectures were used, including: classification and regression tree (CaRT); elastic net logistic regression [23]; support vector machines (SVM); Naïve Bayes; feedforward artificial neural network (NN) with five hidden layers; k-nearest neighbours (KNN); random forest; eXtreme Gradient Boosting (XGBoost).

The best performing algorithm resulting from each architecture was evaluated using AUC, accuracy, sensitivity, specificity, F1 score, negative predictive value (NPV), and positive predictive value (PPV). AUC can be understood as the probability that a randomly chosen non-frail patient will have a score lower than a randomly chosen frail patient. AUC is constructed by plotting the achieved sensitivity and specificity of the classifier at every possible decision threshold level, and measuring the area under the curve. AUC ranges from 0 to 1, with 0.5 being no better than random guessing, and 1 being a perfect classifier. Receiver-operating characteristic (ROC) curves were also constructed and assessed. Although the default decision threshold for binary classification is 0.5, this threshold can also be moved along the ROC curve to account for imbalances in the training data [24], or to maximize both sensitivity and specificity (defined as the point on each curve closest to the upper left corner).

This study followed the RECORD (Reporting of studies Conducted using Observational Routinely-collected health Data) statement [25], with the associated checklist available in the Appendix II.

## Sensitivity analyses

As the original CFS is a 9 point ordinal scale, the assigned CFS scores were dichotomized to reduce the task to a binary classification problem. A scoping review on the usage of CFS in research identified that the majority of studies used a cut-off of 5 and above to define frailty, while fewer studies used a cut-off of 4 and above, and 2 studies used a cut-off of 6 and above [26]. A CFS score of 5 labelled "mildly frail" is also the first time the term "frail" appears in the corresponding label for each frailty score. We will also use cut-off scores of 4 and 6 for sensitivity analyses. By using a cut-off of 4, patients who were identified as 'vulnerable' in the CFS are now considered to be frail, whom were previously labelled to be non-frail. This increased the number of frail patients from the original training set from 724 to 1362, changing the proportion of frail patients from 18.9% to 35.6%. A cut-off of 6 would consider patients were identified as 'mildly frail' to be non-frail, whom were previously labelled frail. This increases the imbalance in the training dataset, as the number of frail patients were reduced from 724 to 358, changing the proportion of frail patients from 18.9% to 9.3%.

All analyses were performed in R version 4.0.4, where the packages 'caret' and 'h2o' were used for model building [27, 28]. SMOTE was implemented using Python 3 through R using the package 'reticulate' [29].

# Results

Of the 5,466 patients sampled, the median age was 74 years (IQR: 11), with 50% of the sample falling between 69 and 80 years of age. The sample had more females than males, with 44% (n = 2,425) of the sample being males. There was 13.4% (n = 732) of the sample that had no known chronic conditions as detected by CPCSSN's validated chronic condition case detection algorithms; of those with known chronic conditions, the most common chronic condition was hypertension (76.35%). The estimated prevalence of frailty among seniors aged 65 and older in this sample of CPCSSN patients was 18.4%.

Compared with non-frail patients (n = 4,460), frail patients (n = 1,006) were statistically significantly likely to be older, female, and less likely to have no known chronic conditions

Table 1: Features used for machine learning

| Features | n | Data type |
|---|---|---|
| Patient age | 1 | Numeric |
| Patient sex | 1 | Binary |
| Patient Diagnoses Received in Last 2 Years (ICD-9 Codes) | 13 | Numeric |
| CPCSSN's Detection of 6 Chronic Conditions | 6 | Binary |
| Medications Prescribed in Last 2 Years | 39 | Numeric |
| Patient Biometrics | 7 | Numeric |
| Province | 1 | Categorical |
| Missing Medication Indicator | 1 | Binary |
| Missing Height, Weight and BMI Indicators | 3 | Binary |
| Missing Chronic Conditions Indicator | 1 | Binary |
| Missing Patient Diagnoses Indicator | 1 | Binary |
| Missing Blood Pressure Indicator | 1 | Binary |
| **Total** | **75** | |

Table 2: Cohort demographics

| | All (n = 5,466) | Frail (n = 1,006) | Not frail (n = 4,460) | p-value |
|---|---|---|---|---|
| Age (Median, [Q1-Q3]) | 74 [69–80] | 81 [74–88] | 72 [68–78] | <0.001† |
| Sex (% Male) | 2,425 (44.4%) | 348 (34.6%) | 2,077(46.6%) | <0.001 |
| No Known Chronic Conditions | 732 (13.4%) | 52 (5.2%) | 680 (15.2%) | <0.001 |
| COPD* | 534 (11.3%) | 382 (10.1%) | 152 (15.9%) | <0.001 |
| Dementia* | 449 (9.5%) | 238 (24.9%) | 211 (5.6%) | <0.001 |
| Depression*1,155 (24.4%) | 316 (33.1%) | 839 (22.2%) | <0.001 | |
| Diabetes Mellitus* | 1,866 (39.4%) | 374 (39.2%) | 1,492 (39.5%) | 0.909 |
| Epilepsy* | 94 (2.0%) | 24 (2.5%) | 70 (1.9%) | 0.237 |
| Hypertension* | 3,614 (76.35) | 760 (79.7%) | 2,854(75.5%) | 0.008 |
| Osteoarthritis* | 2,187 (46.2%) | 439 (46.2%) | 1,748 (46.2%) | 0.929 |
| Mean BMI (Median [Q1–Q3]) | 28.5 [25.31–32.49] | 28.34 [24.52–33.17] | 28.50 [25.40–32.40] | 0.501† |
| Missing BMI | 1,735 (45.3%) | 436 (60.2%) | 1,299 (41.9%) | <0.001 |
| Mean Height (centimetres) (Median [Q1–Q3]) | 165.00 [157.47–173.15] | 160.00 [152.81–168.50] | 165.80 [158.15–174.00] | <0.001† |
| Missing Height (centimetres) | 1761 (46.0%) | 443 (61.2%) | 1318 (42.5%) | <0.001 |
| Mean Weight (kg) (Median [Q1–Q3]) | 79.60 [67.39–92.60] | 75.19 [64.21–90.00] | 80.32 [68.40–93.00] | <0.001† |
| Missing Weight (kg) | 1,317 (34.4%) | 302 (42.7%) | 1,015 (32.7%) | <0.001 |
| Missing Systolic Blood Pressure Measurement | 611 (16.0%) | 111 (15.3%) | 500 (16.1%) | 0.645 |
| Mean Systolic Blood Pressure (Median [Q1–Q3]) | 132.62 [124.50–141.28] | 133.00 [125.33–141.67] | 133.61 [123.95–142.00] | 0.546† |
| Number of Clinic Visits In Most Recent Calendar Year (Median [Q1–Q3]) | 5 [3–9] | 7 [4–11] | 5 [3–9] | <0.001† |
| Missing Clinic Visits | 296 (17.1%) | 35 (4.8%) | 261 (8.4%) | 0.002 |
| Number of Unique Medications Prescribed In Last 2 Years (Median [Q1–Q3]) | 6 [3–10] | 5 [3–9] | 7 [4–11] | <0.001† |
| Missing Medications | 249 (6.5%) | 27 (3.7%) | 222 (7.2%) | 0.001 |

*Proportions of those who has at least one known chronic condition.
†Tested using the Krusal-Wallis test.

as identified by the seven validated CPCSSN case detection algorithms. Of those with at least one chronic condition, frail patients were more likely to have chronic obstructive pulmonary disease (COPD), dementia, depression, and hypertension. Frail patients were also statistically more likely to have a higher number of clinic visits in the most recent calendar year of when their frailty score was given, with a median of seven visits. The proportion of missingness was also unevenly distributed across frailty, with frail patients statistically significantly less likely to have missing BMI, height, weight, clinic visitations and medications.

Figure 1 compares the ROC curves for the models trained using the original imbalanced dataset, and a cut-off of 5 and above as frail.

Figure 1: Comparison of ROC curves for final models trained on original dataset

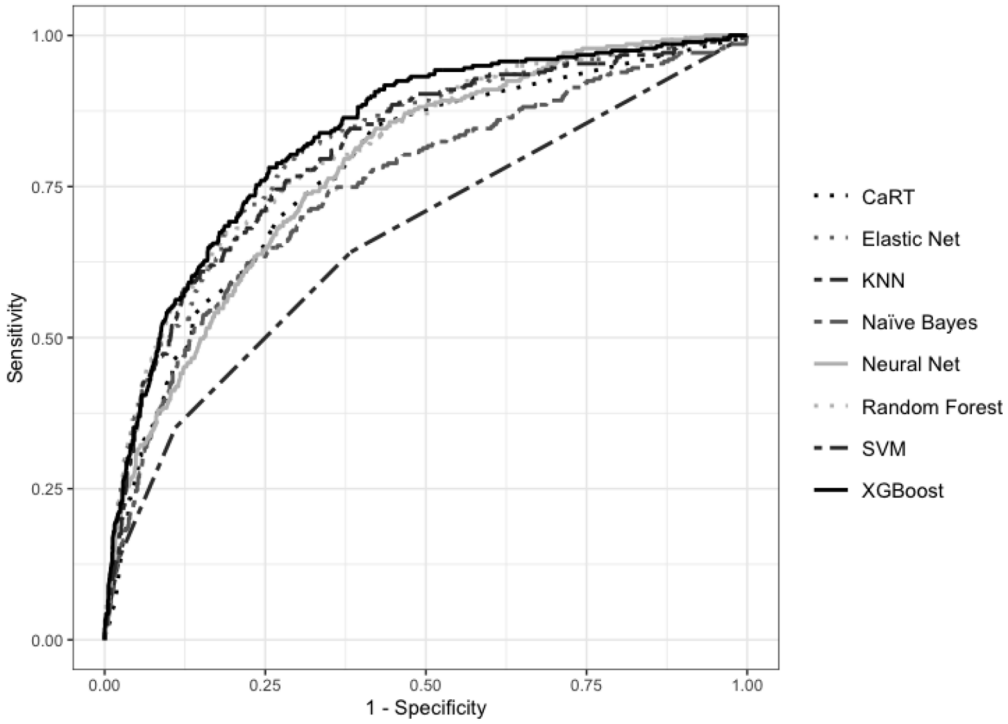


Table 3a: Performance metrics of models trained on original data using hold-out test set

| Model | AUC | Accuracy | F1 | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|---|
| Elastic Net Logistic Regression | 81.58% | 85.42%* | 46.05% | 36.56% | 95.44% | 62.20% | 88.00% |
| SVM | 80.75% | 85.23% | 49.16%* | 41.94% | 94.12% | 59.39% | 88.77% |
| KNN | 66.48% | 83.40% | 21.84% | 13.62% | 97.72%* | 55.07% | 84.65% |
| Naïve Bayes | 74.72% | 70.23% | 43.52% | 67.38%* | 70.81% | 32.14% | 91.37%* |
| CaRT | 77.56% | 82.18% | 44.70% | 42.29% | 90.37% | 47.39% | 88.42% |
| Random Forest | 81.03% | 85.11% | 47.64% | 39.79% | 94.41% | 59.36% | 88.43% |
| XGBoost | 83.18%* | 84.87% | 47.68% | 40.50% | 93.97% | 57.95% | 88.50% |
| Feedforward NN | 78.20% | 84.87% | 35.32% | 24.37% | 97.28% | 64.76%* | 86.25% |

*Highest value achieved for each metric.

Table 3a shows the performance of the 8 supervised machine learning models using the default threshold of 0.5. All models were able to achieve an AUC of over 65%, ranging from 66.48% (KNN) to 83.18% (XGBoost). Sensitivity ranged from 13.62% (KNN) to 67.38% (Naïve Bayes). Specificity ranged from 70.81% (Naïve Bayes) to 97.72% (KNN). PPV ranged from 31.14% (Naïve Bayes) to 64.76% (neural network). NPV ranged from 84.65% (KNN) to 91.37% (Naïve Bayes).

Table 3b: Sensitivity and specificity of models trained on original data using best threshold

| Model | Sensitivity | Specificity | Threshold |
|---|---|---|---|
| Elastic Net Logistic Regression | 77.78% | 72.72% | 0.4730 |
| SVM | 74.55% | 73.38% | 0.1889 |
| KNN | 64.16% | 61.69% | 0.1000 |
| Naïve Bayes | 70.97% | 68.60% | 0.2777 |
| CaRT | 69.89% | 72.79% | 0.1228 |
| Random Forest | 75.27% | 71.99% | 0.3104 |
| XGBoost | 78.14%* | 74.41%* | 0.1851 |
| Feedforward NN | 73.84% | 68.82% | 0.2712 |

*Highest value achieved for each metric.

Table 3b shows the maximum combined sensitivity and specificity that can be achieved by using the most optimal thresholds determined using ROC curves. An XGBoost model achieved the best performance using a threshold of 0.1851, where sensitivity was 78.14% and 74.41%.

Figure 2 compares the ROC curves for the models trained using the balanced dataset created using SMOTE, and a cut-off score of 5 and above as frail.

Table 4a shows the performance of the eight supervised machine learning models trained using the balanced dataset created by SMOTE, and where the default threshold of 0.5 was used. AUC ranged from 65.37% (KNN) to 80.53% (XGBoost). Sensitivity ranged from 30.47% (KNN) to 67.38% (elastic net logistic regression). Specificity ranged from 72.06% (Naïve Bayes) to 93.80% (Random Forest). PPV ranged from 31.53% (Naïve Bayes) to 55.38% (Random Forest). NPV

Figure 2: ROC Curves of models trained on balanced data

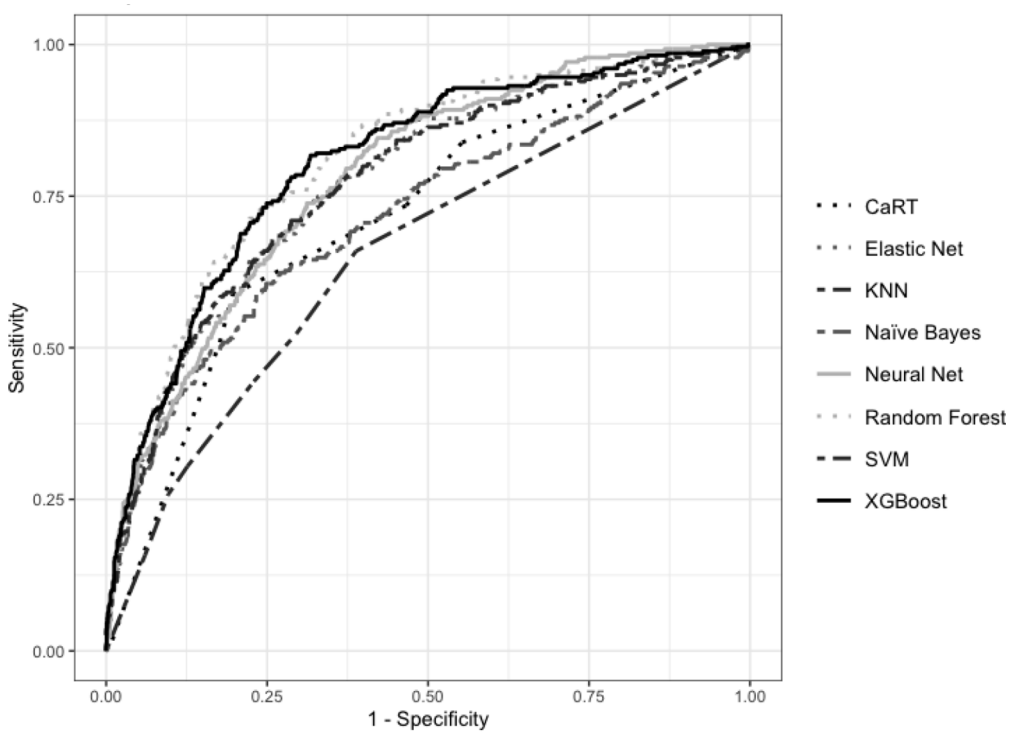


Table 4a: Performance metrics of models trained on balanced data using default threshold

| Model | AUC | Accuracy | F1 | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|---|
| Elastic Net Logistic Regression | 77.21% | 72.79% | 45.74% | 67.38% | 73.90% | 34.62% | 91.70% |
| SVM | 77.26% | 73.89% | 46.10% | 65.59%* | 75.59% | 35.53% | 91.46%* |
| KNN | 65.37% | 77.67% | 31.72% | 30.47% | 87.35% | 33.07% | 85.96% |
| Naïve Bayes | 71.70% | 70.47% | 52.36%* | 62.72% | 72.06% | 31.53% | 90.41% |
| CaRT | 71.27% | 76.69% | 46.20% | 58.78% | 80.37% | 38.05% | 90.48% |
| Random Forest | 80.90%* | 84.20%* | 44.30% | 36.92% | 93.80%* | 55.38%* | 87.89% |
| XGBoost | 80.53% | 83.83% | 44.44% | 37.99% | 93.24% | 53.54% | 87.99% |
| Feedforward NN | 77.76% | 83.28% | 38.01% | 30.11% | 94.19% | 51.53% | 86.79% |

Table 4b: Sensitivity and specificity of models trained on balanced data using best threshold

| Model | Sensitivity | Specificity | Threshold |
|---|---|---|---|
| Elastic Net Logistic Regression | 74.19% | 67.79% | 0.4019 |
| SVM | 70.97% | 71.25% | 0.4429 |
| KNN | 65.95% | 61.18% | 0.0833 |
| Naïve Bayes | 62.01% | 73.75% | 0.6216 |
| CaRT | 58.78% | 80.37% | 0.5121 |
| Random Forest | 72.76% | 76.10% | 0.3525 |
| XGBoost | 77.42%* | 71.84%* | 0.2185 |
| Feedforward NN | 66.67% | 77.72% | 0.6636 |

ranged from 85.96% (KNN) to 91.70% (elastic net logistic regression).

Table 4b shows the maximum combined sensitivity and specificity that can be achieved by using the most optimal thresholds determined using ROC curves. XGBoost achieved the best performance using a threshold of 0.2185, where sensitivity was 77.42% and specificity was 71.84%.

Using a cut-off CFS score of 4 and above as frail, an XGBoost model achieved the best performance of sensitivity (77.42%) using a threshold of 0.3385. A CaRT model achieved the best specificity (76.02%) using a threshold of 0.2540. Using a cut-off CFS score of 6 and above, a CaRT model achieved the highest sensitivity (78.85%) using a threshold of 0.0568. An XGBoost model achieved the best specificity (77.77%) using a threshold of 0.0875. The detailed results of the sensitivity analyses are listed in Appendix III.

The final hyperparameters used for all models are listed in Appendix III Table 5.

# Discussion

This is the first study to use pan-Canadian primary care data to create a frailty case definition using machine learning. We observed a frailty prevalence of 18.4% in the data gathered, which is similar to other reported frailty prevalence estimates in seniors over the age of 65 [30]. A collection of eight common supervised machine learning architectures were used for the identification of dichotomized frailty, and performance assessed using the hold-out test set.

XGBoost had the overall best performance across all training datasets, achieving the highest or second highest sensitivity in each. Using the original imbalanced dataset, an XGBoost model was able to achieve great performance, with 78.14% sensitivity and 74.41% specificity using a decision threshold of 0.1851.

The same XGBoost model achieved a sensitivity of 40.50% and a specificity of 93.97% using a decision threshold of 0.5. We can compare these results with what was achieved previously by Williamson et al, where the CPCSSN EMR data used were only from Alberta and decision threshold used was 0.5 [13]. We can see that by using more machine learning models and a larger dataset, sensitivity was able to improve from 28% to 40.50%, and specificity did not suffer a loss with both at 94%.

The balanced dataset created by SMOTE did not result in better performance as compared with the original imbalanced dataset. One explanation may be that no undersampling was performed. While random undersampling used in tandem with SMOTE can lead to increased classification performance [22], we elected not to undersample the proportion of non-frail patients to preserve information. This lack of undersampling increased the number of synthetic oversampling required to reach a balance between the number of frail and non-frail patients. If the feature space for frail patients and non-frail patients had areas of overlap, oversampling the frail patients may have introduced patients whose label may in actuality be non-frail, inadvertently introducing false positive samples to the training dataset. As Table 2 shows, there was no significant statistical difference in proportion between frail and non-frail patients for diabetes mellitus, epilepsy, osteoarthritis. The distributions of mean BMI and mean systolic blood pressure also had sufficient overlap between that of frail and non-frail patients to not be statistically significantly different.

The sensitivity analyses using two alternative cut-offs for the binary classification of frailty based on the CFS scores resulted in similar performances when the decision threshold was determined using the ROC curves, where the highest sensitivities ranged from 76.37% to 78.85%, and the highest specificities ranged from 76.02% to 77.7%. Changing the cut-off also changed the sample size in each class, affecting the level of imbalance present. When a cut-off of 4 and above was used to identify frailty, the number of frail patients increased. However more noise was potentially introduced as now the sample with the 'frail' label had ranged from patients that were rated 'vulnerable' to those who were rated as 'terminally ill'. Conversely, when a cut-off of 6 and above was used the number of frail patients decreased, resulting in a more severe imbalance and an increase in the noise in the features of the non-frail group.

As there was no significant classification difference between any of the four training datasets, we propose that the best model to use for the identification of frail patients in EMR data is the XGBoost model trained using the original data, with frailty defined using the standard CFS score cut-off of 5 and above. This model is readily deployable, inexpensive, and could be used for public health surveillance and frailty research. As the features used in the model were based on routinely collected structured primary care EMR data, this model could also be easily tested and used in other primary care EMRs.

While our goal was to maximize sensitivity and specificity in tandem, it's also possible to change the decision threshold to other points on the ROC curve that maximize sensitivity at the expense of specificity (and vice versa). For example, the XGBoost model trained on the original imbalanced data with a decision threshold of 0.5 has a sensitivity of 40.50% and a specificity of 93.97%. This model would classify a relatively low number of false positives, and can be used to rule-in frail patients, as patients classified as 'frail' has a high certainty of actually being frail. A model with high specificity could be used for studies assessing the efficacy of interventions for reducing existing level of frailty. These studies may find inconclusive results if both non-frail and frail patients were included, as the level of frailty is unlikely to change for non-frail patients.

The decision threshold may also be moved to achieve a high sensitivity and low specificity, which would result in a low number of false negatives. A model with these characteristics could be used to create frailty screening cohorts, where the goal is to select as many frail patients as possible at the expense of having some false positives.

## Limitations

One important limitation of this study is in the assignment of the CFS scores. The application of the CFS to their own patients may have varied between physicians, and as each patient received only one CFS score, we were also not able to assess inter-rater reliability. Previous research on the inter-rater reliability of the CFS in an emergency care setting showed a kappa of 0.9 between emergency department nurses and emergency department physicians [31]. Another study in an outpatient setting showed an inter-rater reliability of

0.811 for the CFS between physicians [32]. Future on frailty classification may also wish to have physicians rate the same group of patients to assess inter-rater reliability.

Another limitation of clinicians assessing their own patients is that they may have used recall of patient encounters and conversations to assess frailty severity. It was very likely that clinicians used information not recorded in EMR, such as past experiences or intuition in their assessment of the severity of frailty. The classification ability of any model will be hindered if some data used to inform the label was not available. Although this may be an ever present issue in primary care where long-term clinician-patient relationships are common.

Selection bias may have occurred when some physicians had selected a group of their own patients to rate, rather than being provided a list of randomly sampled patients. These physicians may have been more likely to select patients they have seen frequently to better assess their level of frailty. These patients may have higher rates of clinic visitation compared with the average patient in the EMR. Resulting models may be consistently poorer at classifying frailty for patients with few clinic visitations as compared with patients with frequent clinic visitations.

It should be noted that although the task of classification requires a reference-standard label that represents the ground truth, this may not be possible for diseases with unclear or subjective diagnostic criteria. The CFS was created to allow for room for clinical judgement [14], and this flexibility will also introduce wanted variation between patients with the same frailty score on the CFS. However, this variation is undesirable for supervised machine learning. Future research may wish to use multiple raters to assess each senior patient on their level of frailty, and assess the differences between patients who had varied CFS scores versus those who had consistent CFS scores.

While dichotomization is common practice in disease identification, it reduces the amount of information that can be used. Patients who were previously separated by frailty severity are now one common class, where mildly frail patients have the same label as severely frail patients. As we had dichotomized the CFS after the physicians had rated their patients, it is also possible that some physicians would have disagreed with the cut-off of 5 to define frailty. This may have been another source of variation introduced to the data. Future work may wish to keep the original 9 point ordinal scale, or collapsed groups of 4 or 5 levels of frailty to increase the sample size in each category. An alternative approach could be to assess the CFS as a continuous variable, by approximating the underlying distribution to the distribution of the nine classes, then creating decision boundaries for the transformation back to the ordinal CFS to assess performance.

One of the challenges of using EMR data is the lack of standard in how each EMR database may record, process, and store their information [33, 34]. This study combined data from five different regional CPCSSN networks, each one within a unique province in Canada. Each regional CPCSSN network had provided the most recent extraction of their data, which had been cleaned and processed using their own methods. Not all networks provided EMR records containing unstructured data, thus all available data for featurization were reduced to structured data that were collected in all networks. This was a large limitation as while processed data may be more readily used, unstructured free-text notes have been shown to contain diagnostic suspicion that was not coded [35] and potential disease incidence [36]. Kharrazi et al. showed that geriatric syndromes were significantly more likely to be identified using unstructured EMR notes as compared with structured data only [37]. Specifically, the addition of free-text notes processed using natural language processing methods increased the detection rate of geriatric syndromes by a factor of 3.2 times for falls, 18 for malnutrition, 3.4 for walking difficulties, and 455.9 for lack of social support. Future studies could link primary EMR data with other data sources, such as hospitalization or emergency care records, or specialist outpatient clinics EMRs to increase the amount of available data.

## Conclusion

We were able to create a supervised classification model using XGBoost for the identification of frailty with a 78.14% sensitivity and 74.41% specificity using routinely collected primary care EMR data for usage in the Canadian context. This classification model could be used for further research on frail patients within primary care, as well as for public health surveillance.

Neither the use of alternative cut-offs for the definition of frailty nor the use of SMOTE for minority oversampling resulted in a change in classification performance. Future research may consider using physicians to rate the same group of patients to assess for inter-rater reliability, and to supplement primary care EMR data with data from other sources in the healthcare system.

## Funding

This research study was funded by The Canadian Frailty Network. Preceding research resulting in Alberta data used for this current research study was funded by The Canadian Institutes for Health Research (SQS-145182), Michael Smith Foundation for Health Research (#16734) and the Canadian Frailty Network.

## Acknowledgements

We would like to acknowledge the family physicians in Alberta, British Columbia, Manitoba, Ontario, and Nova Scotia that took part in completing the Rockwood Clinical Frailty Scale.

## Conflict of interest

The authors declare no conflicts of interest.

## Ethics

All procedures and analyses were approved by regional network directors' ethics boards, including: University of British Columbia (REB# H18-01341), University of Calgary (REB# 18-1881), University of Manitoba (REB # HS22406 (H2018:486)), McMaster University (REB# 5393), and Dalhousie University (REB #1024172).

## Accessibility of data and code

Similar data are available from CPCSSN upon request. Code is available upon request from authors.

## References


1. Walston J, Hadley EC, Ferrucci L, Guralnik JM, Newman AB, Studenski SA, et al. Research agenda for frailty in older adults: toward a better understanding of physiology and etiology: summary from the American Geriatrics Society/National Institute on Aging Research Conference on Frailty in Older Adults. J Am Geriatr Soc. 2006 Jun;54(6):991–1001. https://doi.org/10.1111/j.1532-5415.2006.00745.x.

2. Yanagawa B, Latter DA, Fedak PWM, Cutrara C, Verma S. The Cost of Frailty in Cardiac Surgery. Can J Cardiol. 2017 Aug 1;33(8):959–60. https://doi.org/10.1016/j.cjca.2017.05.015.

3. Rodrigues MK, Marques A, Lobo DML, Umeda IIK, Oliveira MF. Pre-Frailty Increases the Risk of Adverse Events in Older Patients Undergoing Cardiovascular Surgery. Arq Bras Cardiol. 2017 Oct;109(4):299–306. https://doi.org/10.5935/abc20170131.

4. Chen C-L, Chen C-M, Wang C-Y, Ko P-W, Chen C-H, Hsieh C-P, et al. Frailty is Associated with an Increased Risk of Major Adverse Outcomes in Elderly Patients Following Surgical Treatment of Hip Fracture. Sci Rep. 2019 Dec;9(1):19135. https://doi.org/10.1038/s41598-019-55459-2.

5. Crocker TF, Brown L, Clegg A, Farley K, Franklin M, Simpkins S, et al. Quality of life is substantially worse for community-dwelling older people living with frailty: systematic review and meta-analysis. Qual Life Res. 2019 Aug 1;28(8):2041–56. https://doi.org/10.1007/s11136-019-02149-1.

6. Oldenkamp M, Hagedoorn M, Wittek R, Stolk R, Smidt N. The impact of older person's frailty on the care-related quality of life of their informal caregiver over time: results from the TOPICS-MDS project. Qual Life Res. 2017 Oct 1;26(10):2705–16. https://doi.org/10.1007/s11136-017-1606-5.

7. Frailty Matters [Internet]. Canadian Frailty Network. [cited 2020 Dec 6]. Available from: https://www.cfn-nce.ca/frailty-matters/.

8. Travers J, Romero-Ortuno R, Bailey J, Cooney M-T. Delaying and reversing frailty: a systematic review of primary care interventions. Br J Gen Pract. 2019 Jan 1;69(678):e61–9. https://doi.org/10.3399/bjgp18X700241.

9. Lacas A, Rockwood K. Frailty in primary care: a review of its conceptualization and implications for practice. BMC Med. 2012 Jan 11;10(1):4. https://doi.org/10.1186/1741-7015-10-4.

10. Braithwaite RS, Fiellin D, Justice AC. The Payoff Time. Med Care. 2009 Jun;47(6):610–7. https://doi.org/10.1097/MLR.0b013e31819748d5.

11. Collard RM, Boter H, Schoevers RA, Voshaar RCO. Prevalence of Frailty in Community-Dwelling Older Persons: A Systematic Review. J Am Geriatr Soc. 2012;60(8):1487–92. https://doi.org/10.1111/j.1532-5415.2012.04054.x.

12. Lethebe BC. Using machine learning methods to improve chronic disease case definitions in primary care electronic medical records. 2018 Apr 23 [cited 2019 Oct 18]; Available from: https://prism.ucalgary.ca/handle/1880/106538. https://doi.org/10.11575/PRISM/31824.

13. Williamson T, Aponte-Hao S, Mele B, Lethebe BC, Leduc C, Thandi M, et al. Developing and validating a primary care EMR-based frailty definition using machine learning. Int J Popul Data Sci. 2020 Sep 1;5(1):1344. https://doi.org/10.23889/ijpds.v5i1.1344.

14. Rockwood K, Song X, MacKnight C, Bergman H, Hogan DB, McDowell I, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ. 2005 Aug 30;173(5):489–95. https://doi.org/10.1503/cmaj.050051.

15. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in Older Adults: Evidence for a Phenotype. J Gerontol A Biol Sci Med Sci. 2001 Mar 1;56(3):M146–57. https://doi.org/10.1093/gerona/56.3.m146.

16. Hassler AP, Menasalvas E, García-García FJ, Rodríguez-Mañas L, Holzinger A. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. BMC Med Inform Decis Mak. 2019 Feb 18;19(1):33. https://doi.org/10.1186/s12911-019-0747-6.

17. Clegg A, Bates C, Young J, Ryan R, Nichols L, Ann Teale E, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. Age Ageing. 2016 May 1;45(3):353–60. https://doi.org/10.1093/ageing/afw039.

18. Ambagtsheer RC, Shafiabady N, Dent E, Seiboth C, Beilby J. The application of artificial intelligence (AI) techniques to identify frailty within a residential aged care administrative data set. Int J Med Inf. 2020 Apr 1;136:104094. https://doi.org/10.1016/j.ijmedinf.2020.104094.

19. Garies S, Birtwhistle R, Drummond N, Queenan J, Williamson T. Data Resource Profile: National electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Int J Epidemiol. 2017 Aug 1;46(4):1091–1092f. https://doi.org/10.1093/ije/dyw248.

20. Kadhim-Saleh A, Green M, Williamson T, Hunter D, Birtwhistle R. Validation of the Diagnostic Algorithms for 5 Chronic Conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): A Kingston Practice-based Research Network (PBRN) Report. J Am Board Fam Med. 2013 Mar 1;26(2):159–67. https://doi.org/10.3122/jabfm.2013.02.120183.

21. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: What is it and how does it work? Int J Methods Psychiatr Res. 2011 Mar 1;20(1):40–9. https://doi.org/10.1002/mpr.329.

22. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002 Jun 1;16:321–57. https://doi.org/10.1613/jair.953.

23. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

24. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978 Oct 1;8(4):283–98. https://doi.org/10.1016/s0001-2998(78)80014-2.

25. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement [Internet]. [cited 2020 Dec 6]. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001885. https://doi.org/10.1371/journal.pmed.1001885.

26. Church S, Rogers E, Rockwood K, Theou O. A scoping review of the Clinical Frailty Scale. BMC Geriatr. 2020 Oct 7;20(1):393. https://doi.org/10.1186/s12877-020-01801-7.

27. Kuhn M. caret: Classification and Regression Training. ascl. 2015 May;ascl:1505.003.

28. LeDell E, Gill N, Aiello S, Fu A, Candel A, Click F, et al. h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. R package version. 3.32.1.3. Available from: httsp://CRAN.R-project.org/package=h2o.

29. Ushey K, Allaire JJ, Tang Y. reticulate: Interface to "Python". R package version 1.20. Available from: https://CRAN.R-project.org/package=reticulate.

30. Kehler DS, Ferguson T, Stammers AN, Bohm C, Arora RC, Duhamel TA, et al. Prevalence of frailty in Canadians 18–79 years old in the Canadian Health Measures Survey. BMC Geriatr. 2017 Jan 21;17(1):28. https://doi.org/10.1186/s12877-017-0423-6.

31. Lo AX, Heinemann AW, Gray E, Lindquist LA, Kocherginsky M, Post LA, et al. Inter-rater Reliability of Clinical Frailty Scores for Older Patients in the Emergency Department. Acad Emerg Med. 2021;28(1):110–3. https://doi.org/10.1111/acem.13953.

32. Özsürekci C, BalcıC, Kızılarslanoğlu MC, Çalışkan H, Tuna Doğrul R, Ayçiçek GŞ, et al. An important problem in an aging country: identifying the frailty via 9 Point Clinical Frailty Scale-. Acta Clin Belg. 2020 May 3;75(3):200–4. https://doi.org/10.1080/17843286.2019.1597457.

33. Kumar S, Aldrich K. Overcoming barriers to electronic medical record (EMR) implementation in the US healthcare system: A comparative study. Health Informatics J. 2010 Dec 1;16(4):306–18. https://doi.org/10.1177/1460458210380523.

34. Sachdeva S, Bhalla S. Semantic interoperability in standardized electronic health record databases. J Data Inf Qual. 2012 May 7;3(1):1:1-1:37. https://doi.org/10.1145/2166788.2166789.

35. Ford E, Nicholson A, Koeling R, Tate AR, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol. 2013 Aug 21;13(1):105. https://doi.org/10.1186/1471-2288-13-105.

36. Koeling R, Tate AR, Carroll JA. Automatically estimating the incidence of symptoms recorded in GP free text notes. In: Proceedings of the first international workshop on Managing interoperability and complexity in health systems [Internet]. New York, NY, USA: Association for Computing Machinery; 2011 [cited 2021 Feb 20]. p. 43–50. (MIXHS '11). https://doi.org/10.1145/2064747.2064757.

37. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The Value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification. J Am Geriatr Soc. 2018;66(8):1499–507. https://doi.org/10.1111/jgs.15411.

## Abbreviations

| | |
|---|---|
| CPCSSN: | Canadian Primary Care Sentinel Surveillance Network |
| AUC: | Area Under Receiver Operating Characteristic Curve |
| CaRT: | Classification and Regression Tree |
| CFS: | Clinical Frailty Score |
| COPD: | Chronic Obstructive Pulmonary Disease |
| EMR: | Electronic Medical Record |
| KNN: | K-Nearest Neighbours |
| NPV: | Negative Predictive Value |
| PPV: | Positive Predictive Value |
| ReLU: | Rectified Linear Unit |
| ROC: | Receiver Operating Characteristic |
| SMOTE: | Synthetic Minority Over-sampling Technique |
| SMOTE-NC: | Synthetic Minority Over-sampling Technique-Nominal Continuous |
| SVM: | Support Vector Machines |
| XGBoost: | Extreme Gradient Boosting |

# Appendix I

Figure A1: Clinical frailty scale


Rockwood K, Song X, MacKnight C, Bergman H, Hogan DB, McDowell I, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ. 2005 Aug 30;173(5):489–95.

Appendix II. The RECORD statement – checklist of items, extended from the STROBE statement that should be reported in observational studies using routinely collected health data

| | Item No. | STROBE items | Location in manuscript where items are reported | RECORD items | Location in manuscript where items are reported |
|---|---|---|---|---|---|
| **Title and abstract** | | | | | |
| | 1 | (a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found | | RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included.<br>RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract.<br>RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract. | 1.1 - 1.3 are all reported in abstract (page 1). |
| **Introduction** | | | | | |
| Background rationale | 2 | Explain the scientific background and rationale for the investigation being reported | | | Pages 2 - 3 |
| Objectives | 3 | State specific objectives, including any prespecified hypotheses | | | Page 3 |
| **Methods** | | | | | |
| Study Design | 4 | Present key elements of study design early in the paper | | | Page 4 |
| Setting | 5 | Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection | | | Page 4 |
| Participants | 6 | *(a) Cohort study* - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up<br>*Case-control study* - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls<br>*Cross-sectional study* - Give the eligibility criteria, and the sources and methods of selection of participants<br>*(b) Cohort study* - For matched studies, give matching criteria and number of exposed and unexposed<br>*Case-control study* - For matched studies, give matching criteria and the number of controls per case | | RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided.<br>RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided.<br>RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage. | Pages 4 - 5 |
| Variables | 7 | Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable. | | RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided. | Page 5 |
| Data sources/ measurement | 8 | For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group | | | Page 5 |

Appendix II. Continued

| | Item No. | STROBE items | Location in manuscript where items are reported | RECORD items | Location in manuscript where items are reported |
|---|---|---|---|---|---|
| Bias | 9 | Describe any efforts to address potential sources of bias | | | Page 6 |
| Study size | 10 | Explain how the study size was arrived at | | | Page 5 |
| Quantitative variables | 11 | Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why | | | Page 5 |
| Statistical methods | 12 | (a) Describe all statistical methods, including those used to control for confounding<br>(b) Describe any methods used to examine subgroups and interactions<br>(c) Explain how missing data were addressed<br>(d) *Cohort study* - If applicable, explain how loss to follow-up was addressed<br>*Case-control study* - If applicable, explain how matching of cases and controls was addressed<br>*Cross-sectional study* - If applicable, describe analytical methods taking account of sampling strategy<br>(e) Describe any sensitivity analyses | | | Pages 5 - 7 |
| Data access and cleaning methods | | .. | | RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population.<br>RECORD 12.2: Authors should provide information on the data cleaning methods used in the study. | Pages 4 - 5 |
| Linkage | | .. | | RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided. | N/A |
| **Results** | | | | | |
| Participants | 13 | (a) Report the numbers of individuals at each stage of the study (*e.g.*, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed)<br>(b) Give reasons for non-participation at each stage.<br>(c) Consider use of a flow diagram | | RECORD 13.1: Describe in detail the selection of the persons included in the study (*i.e.*, study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram. | Pages 4 - 6 |
| Descriptive data | 14 | (a) Give characteristics of study participants (*e.g.*, demographic, clinical, social) and information on exposures and potential confounders<br>(b) Indicate the number of participants with missing data for each variable of interest<br>(c) *Cohort study* - summarise follow-up time (*e.g.*, average and total amount) | | | Pages 7 - 9 |

Appendix II. Continued

| | Item No. | STROBE items | Location in manuscript where items are reported | RECORD items | Location in manuscript where items are reported |
|---|---|---|---|---|---|
| Outcome data | 15 | *Cohort study* - Report numbers of outcome events or summary measures over time<br>*Case-control study* - Report numbers in each exposure category, or summary measures of exposure<br>*Cross-sectional study* - Report numbers of outcome events or summary measures | | | Page 8 |
| fMain results | 16 | (a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included<br>(b) Report category boundaries when continuous variables were categorized<br>(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period | | | Page 9 - 10 |
| Other analyses | 17 | Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses | | | N/A |
| **Discussion** | | | | | |
| Key results | 18 | Summarise key results with reference to study objectives | | | Pages 10 - 11 |
| Limitations | 19 | Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias | | RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported. | Pages 10 - 13 |
| Interpretation | 20 | Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence | | | Page 13 |
| Generalisability | 21 | Discuss the generalisability (external validity) of the study results | | | Page 13 |
| **Other Information** | | | | | |
| Funding | 22 | Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based | | | Page 14 |
| Accessibility of protocol, raw data, and programming code | | .. | | RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code. | Page 14 |

# Appendix III

Figure 1: ROC Curves of models trained on original dataset (cut-off of 4)

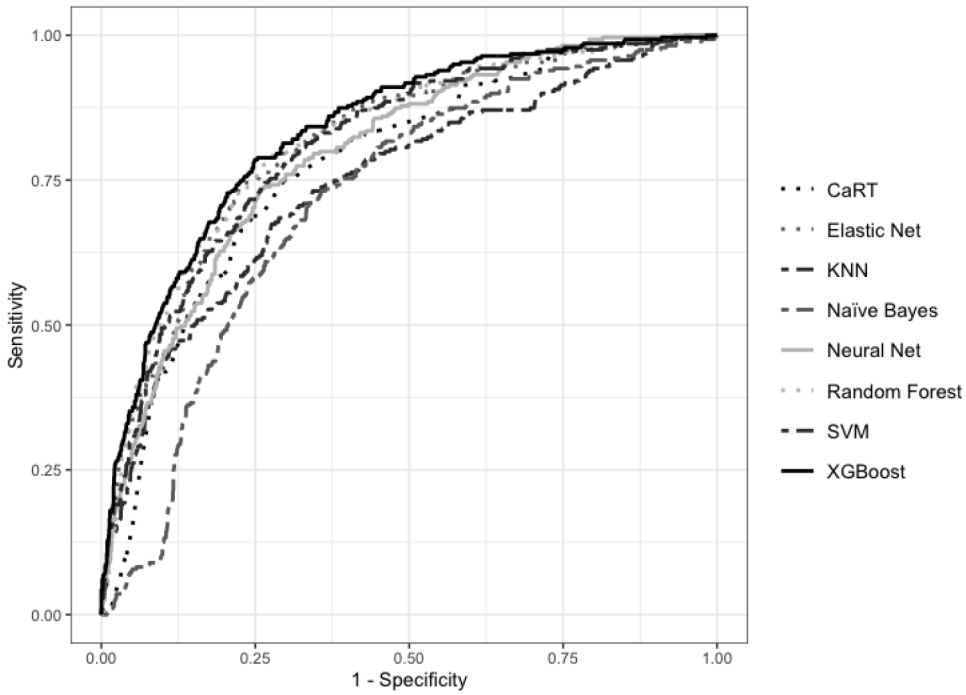


Table 1: Performance metrics of models trained on original data using default threshold (cut-off of 4)

| Model | AUC | Accuracy | F1 | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|---|
| Elastic Net Logistic Regression | 81.43% | 77.30% | 51.56% | 70.97% | 78.60% | 40.49% | 92.96% |
| SVM | 79.01% | 73.52% | 60.62% | 57.19% | 82.56% | 64.48% | 77.70% |
| KNN | 73.46% | 72.18% | 45.71% | 32.88% | 93.93% | 75.00% | 71.66% |
| Naïve Bayes | 68.48% | 66.20% | 42.65% | 73.84% | 64.63% | 29.99% | 92.33% |
| CaRT | 75.67% | 82.98% | 46.66% | 68.82% | 74.12% | 35.29% | 92.05% |
| Random Forest | 79.36% | 75.41% | 63.50% | 58.39% | 85.88% | 69.59% | 78.85% |
| XGBoost | 81.91% | 76.08% | 53.06% | 73.12% | 78.97% | 41.63% | 93.47% |
| Feedforward NN | 79.56% | 81.03% | 47.02% | 49.46% | 87.50% | 44.81% | 89.41% |

Table 2: Sensitivity and specificity of models trained on original data using best threshold – cut-off of 4

| Model | Sensitivity | Specificity | Threshold |
|---|---|---|---|
| Elastic Net Logistic Regression | 74.55% | 77.06% | 0.4787 |
| SVM | 71.75% | 72.51% | 0.3754 |
| KNN | 65.92% | 70.05% | 0.2914 |
| Naïve Bayes | 61.30% | 70.33% | 0.0000 |
| CaRT | 64.90% | 76.02%* | 0.2540 |
| Random Forest | 72.26% | 73.36% | 0.4070 |
| XGBoost | 76.37%* | 71.75% | 0.3385 |
| Feedforward NN | 73.48% | 74.19% | 0.5534 |

Figure 2: ROC Curves of models trained on original dataset (cut-off of 6)

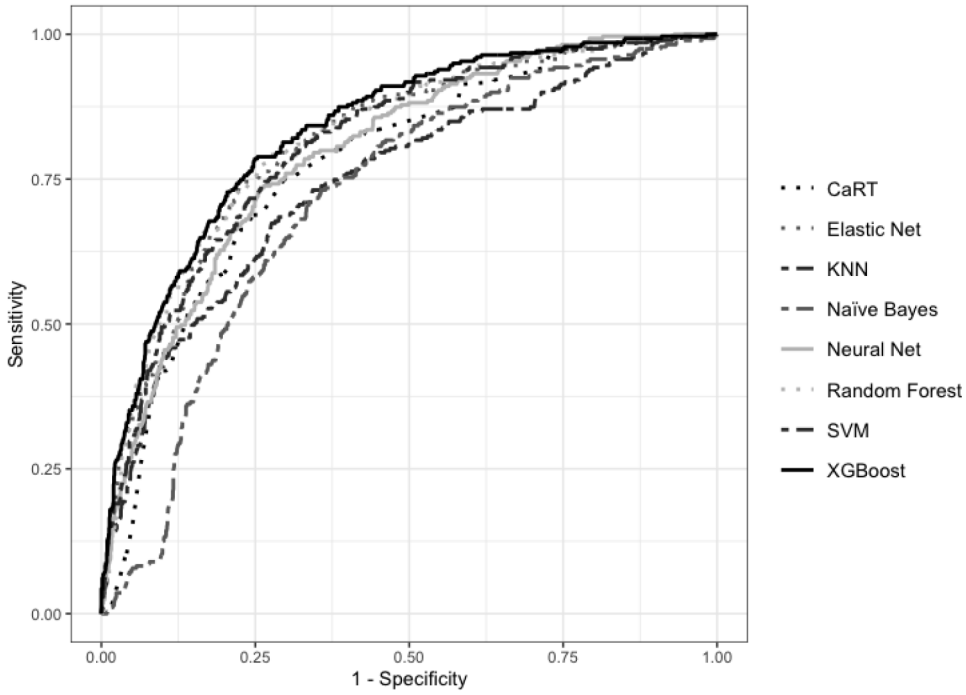


Table 3: Performance metrics of models trained on original data using default threshold (cut-off of 6)

| Model | AUC | Accuracy | F1 | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|---|---|
| Elastic Net Logistic Regression | 80.83% | 84.32% | 23.28% | 13.98% | 98.75% | 69.64% | 84.84% |
| SVM | 71.49% | 78.95% | 39.79% | 40.86% | 86.76% | 38.78% | 87.73% |
| KNN | 73.06% | 64.92% | 4.64% | 2.40% | 99.53% | 73.68% | 64.81% |
| Naïve Bayes | 72.21% | 71.63% | 43.64% | 64.52% | 73.09% | 32.97% | 90.94% |
| CaRT | 78.81% | 90.85% | 32.13% | 32.67% | 92.88% | 31.61% | 93.19% |
| Random Forest | 78.86% | 91.28% | 11.18% | 6.00% | 99.87% | 81.82% | 91.34% |
| XGBoost | 83.70% | 91.64% | 27.91% | 17.20% | 98.75% | 73.85% | 92.95% |
| Feedforward NN | 75.04% | 90.97% | 11.90% | 6.67% | 99.46% | 55.56% | 91.36% |

Table 4: Sensitivity and specificity of models trained on original data using best threshold – cut-off of 6

| Model | Sensitivity | Specificity | Threshold |
|---|---|---|---|
| Elastic Net Logistic Regression | 73.12% | 76.47% | 0.0806 |
| SVM | 67.33% | 72.87% | 0.1277 |
| KNN | 67.33% | 66.62% | 0.0255 |
| Naïve Bayes | 59.33% | 80.05% | 0.9973 |
| CaRT | 78.85%* | 71.99% | 0.0568 |
| Random Forest | 70.67% | 70.92% | 0.2086 |
| XGBoost | 76.00% | 77.77%* | 0.0875 |
| Feedforward NN | 74.19% | 64.34% | 0.0000 |

Table 5: Hyperparameters used for final models

| Model | Original imbalanced data using cut-off of 5 | SMOTE balanced data using cut-off of 5 | Original imbalanced data using cut-off of 4 | Original imbalanced data using cut-off of 6 |
|---|---|---|---|---|
| **Elastic Net Logistic Regression** | alpha = 0.5318833, lambda = a0.005369339 | alpha = 0.1764004, lambda = 0.002016792 | alpha = 0.5600862, lambda = 7.090597 | alpha = 0.1, lambda = 0.01925033 |
| **SVM** | polynomial kernel, degree = 3, scale = 0.004422882, C = 0.1504941 | radial kernel, sigma = 0.02996594, C = 170.478 | degree = 2, scale = 0.0005473211, C = 267.0139 | linear kernel, C = 181.4091 |
| **KNN** | kmax = 55, distance = 0.2262503, kernel = triweight | kernel = rank, distance = 1, kmax = 500 | kmax = 105, distance = 1.644928, kernel = cos | kmax = 1043, distance = 0.9733469, kernel = triweight |
| **Naïve Bayes** | fL = 0, usekernel = True, adjust = 1 | fL = 0.1, no kernel usage, adjust = 0.5 | fL = 0, usekernel= T, adjust = 1 | fL = 0, usekernel = F, adjust = 1 |
| **CaRT** | cp = 0.0002762431 | cp = 0.009829198 | cp = 0.00201909 | cp = 0 |
| **Random Forest** | mtry = 11, splitrule = gini, min.node.size = 9 | mtry= 3, splitrule = gini, min.node.size = 2 | mtry = 11, splitrule = gini, min.node.size = 9 | mtry = 12 |
| **XGBoost** | nrounds = 971, max_adepth = 2, eta = 0.2322766, gamma = 5.086296, colsample_bytree = 0.5705734, min_child_weight = 18, subsample = 0.9047023 | nrounds = 365, max_depth = 2, eta = 0.2394084, gamma = 9.56787, colsample_bytree = 0.3579414, min_child_weight = 5, subsample = 0.6451248 | nrounds = 707, max_depth = 6, eta = 0.06909712, gamma = 6.766357, colsample_bytree = 0.3710754, min_child_wight = 1, subsample = 0.7310282 | nrounds = 714, max_depth = 7, eta = 0.06228869, gamma = 7.277172, colsample_bytree = 0.3480463, min_child_weight = 15, subsample = 0.5177022 |
| **Feedforward NN** | epochs = 500, hidden = c(100, 100, 100, 100, 100), activation = 'MaxoutWithDropOut', dropout = 50%, loss =CrossEntropy | epochs = 500, hidden = c(100, 100, 100, 100, 100), activation = 'MaxoutWithDropOut', dropout = 50%, loss =CrossEntropy | epochs = 500, hidden = c(100, 100, 100, 100, 100), activation = 'MaxoutWithDropOut', dropout = 50%, loss =CrossEntropy | epochs = 500, hidden = c(100, 100, 100, 100, 100), activation = 'MaxoutWithDropOut', dropout = 50%, loss =CrossEntropy |