Frail People in LABLand: Development of an Easy-to-Use Machine Learning Model to Identify Frail People in Hospitals Based on Laboratory Data

Julian GUTHEILa,1, Philip STAMPFERa, Diether KRAMERb,c, Manuel WECHSELBERGERb, Sai Pavan Kumar VEERANKIb,c, Michael SCHREMPFb,c, Peter MRAKc, Martina AUBELa and Franz FEICHTNERa

a Joanneum Research Forschungsgesellschaft mbH, Graz, Austria
b PH Predicting Health GmbH, Graz, Austria
c Steiermärkische Krankenanstaltengesellschaft mbH, Graz, Austria

Abstract. Background: Frail individuals are very vulnerable to stressors, which often lead to adverse outcomes. To ensure an adequate therapy, a holistic diagnostic approach is needed which is provided in geriatric wards. It is important to identify frail individuals outside the geriatric ward as well to ensure that they also benefit from the holistic approach. Objectives: The goal of this study was to develop a machine learning model to identify frail individuals in hospitals. The model should be applicable without additional effort, quickly and in many different places in the healthcare system. Methods: We used Gradient Boosting Decision Trees (GBDT) to predict a frailty target derived from a gold standard assessment. The used features were laboratory values, age and sex. We also identified the most important features. Results: The best GBDT achieved an AUROC of 0.696. The most important laboratory values are urea, creatinine, granulocytes, chloride and calcium. Conclusion: The model performance is acceptable, but insufficient for clinical use. Additional laboratory values or the laboratory history could improve the performance.

Keywords. Frailty, Machine Learning, Hospital Laboratory, Diagnostic Screening Programs.

1. Introduction

Frail individuals are vulnerable to stressors and often lack the resources to cope with such stressors [1–3]. A stressor in frail individuals is often followed by adverse outcomes [2, 4]. These include e.g. a significant reduction in independence, delirium, falls or earlier mortality [2, 4, 5]. Frailty is understood to be a dynamic condition, which can be reduced by appropriate interventions [6–8]. An important characteristic of frail individuals is the impairment of several physiological systems of the body [1–3, 8, 9], which complicates identifying an appropriate therapy as complex interrelationships need to be taken into account. Current healthcare systems are often fragmented and focus on one disease at a

1 Corresponding Author: Julian Gutheil, Joanneum Research Forschungsgesellschaft mbH, Graz, Austria, E-Mail: healthca@joanneum.at
time, neglecting comorbidities [10, 11]. However, treatment of frail individuals requires a holistic approach, which is often initiated by a Comprehensive Geriatric Assessment (CGA), followed by an appropriate interdisciplinary therapy [10, 12]. The CGA is used to assess relevant dimensions (physical, social, and psychological) of the patient status in an objective and structured way. Several studies have shown that frail individuals benefit from a CGA [12–14]. As frail individuals are not exclusively found in geriatric wards, they have to be identified in order to ensure an adequate therapy for these individuals.

1.1. Frailty

There are two main models to identify frail people [2]: the Phenotype Model [4] and the Cumulative Deficit Model [15, 16]. The Phenotype Model consists of five variables that typically occur in frail individuals: exhaustion, unintentional weight loss, low physical activity, slow walking speed, and weak grip strength. An individual is classified as frail if three or more variables exceed a certain threshold. The Frailty Index [15] is based on the Cumulative Deficit Model. Using a list of 70 clinical deficits, the deficits for each individual are summed. The more deficits an individual has accumulated, the more likely this individual is frail.

1.2. Frailty Assessments

CGA acts as the gold standard to identify frail individuals, but it is also very time-consuming and requires geriatric expertise [2]. Consequently, the use of CGA outside of geriatrics is not practicable. According to Bruyère et al. [17], the Gait Speed Test, the Clinical Frailty Scale [18], the Short Physical Performance Battery, the Frailty Phenotype, and the Frailty Index are mainly used in clinical practice to identify frailty. According to Dent et al. [19, 20], the most commonly used frailty assessments in clinical practice and research are the Frailty Phenotype and the Frailty Index.

Frailty assessments for clinical practice should be quick to perform, easy to use, inexpensive, and reliable [21]. Major barriers in everyday clinical practice are time pressure or lack of time [22]. Assessments currently used in practice are time-consuming and therefore unsuitable for large-scale screening.

According to Kim, there are three categories of electronic frailty assessments that have been developed using healthcare data, which can be automated [23]: (a) Clinical Knowledge-Driven Selection: usage of expertise to gather certain characteristics and diagnoses that are associated with frailty. The frailty score is then derived from the presence or absence of these characteristics and diagnoses. (b) Data-Driven Selection without a Reference Standard: Usage of electronic health records (EHRs) for a cluster analysis. Then an attempt is made to identify a frailty cluster (like e.g. the Hospital Frailty Risk Score (HFRS) [24]). (c) Data-Driven Selection with a Reference Standard: Usage of EHRs to train a machine learning (ML) model. Compared to category (b), the EHRs include a frailty target (e.g. Frailty Phenotype), which will be predicted using the ML model. This approach is also used in this paper.

Our data includes a frailty target assessed by a CGA, which provides a high quality target that, to our knowledge, has not yet been used for a category (c) model. According to Kim, the models developed to date use constructs related to frailty (such as disability) or the Frailty Phenotype, as well as a Deficit Accumulation Frailty Index [23].
1.3. Frailty and Laboratory Values

We used laboratory values as predictors for the frailty model, as they are routinely collected in healthcare facilities, represent objective parameters and are well comparable across healthcare facilities. In category (c) models listed by Kim, laboratory values have not yet been used as predictors [23]. Hao et al., Wang et al., and Ellis et al. [25–27] have developed a Deficit Accumulation Frailty Index using laboratory data, without using a ML approach. Laboratory data outside the normal range were considered a deficit, and the number of deficits subsequently yields the frailty score.

1.4. Objective

The goal of this study was to develop a ML model to identify frail individuals in hospitals using existing laboratory results. The model should be applicable without additional effort, quickly and in many different places in the healthcare system. Furthermore, it should provide geriatric expertise outside of geriatric wards.

2. Methods

2.1. Dataset

We used retrospective real world data from the Benchmarking and Reporting System (BARS) used in Austrian acute geriatrics, which was collected using a standardized form for each hospital stay of each patient. Data of five acute geriatric care facilities was merged with corresponding data from the hospital information system (HIS) of Steiermärkische Krankenanstaltengesellschaft m. b. H. (KAGes). HIS data was extracted for individuals admitted to an acute geriatric care facility from August 2008 to April 2021. In total, the dataset included 14,703 cases. Extracted HIS data contained the variables “sex”, “age”, “mortality during stay”, HFRS and the 60 most common laboratory values collected from 30 days before admission until 8 pm on the day of admission.

To develop a generalizable model, we included only individuals with a complete blood count. In addition, we only considered the 35 laboratory values that were present in at least 80% of cases. On the one hand, we wanted to consider only laboratory values that are routinely collected. On the other hand, we wanted to avoid that the mere presence of a blood value provides information content. Geriatricians might recognize an individual as frail and therefore request a specific blood test.

For model development, only cases without missing data in the variables “sex”, “age” and the frailty target were included. We used only one case per person to avoid dependencies between training and test set. The frailty target, the variable “delirium during stay” and the variable “fall during stay” originated from BARS data. The frailty target was labeled by geriatricians, based on the CGA. The variables “delirium during stay”, “fall during stay”, “mortality during stay” and HFRS originated from HIS data and were used to validate the trained model. The variables “sex”, “age” and the laboratory values were used as model features.

The final dataset consisted of 8,506 individuals (69.6% female/30.4% male). The average age was 77.5±9.3 years. As a result of the CGA, 19.4% were classified as frail.
2.2. Machine Learning Methods

We used Gradient Boosting Decision Trees (GBDT), such as XGBoost [28], LightGBM [29], and CatBoost [30] for model development, as they represent state-of-the-art ML models for tabular data [31–34]. We also used a logistic regression to estimate the required model complexity for the frailty patterns.

2.3. Model Development

We used a nested Cross Validation (CV): the dataset was randomly divided into five equal-sized folds. 5-fold CV was applied to estimate the generalization error. Additionally, 4-fold CV was used within the training folds for Bayesian Hyperparameter Optimization (HPO). The HPO identifies the hyperparameters that provide the largest averaged Area Under the Receiver Operating Characteristic (AUROC) across each of the four validation folds. The hyperparameter search space was adopted from Schäfl et al. [35]. Bayesian HPO was performed using the hyperopt package, as by Shwartz-Ziv et al. [31]. For XGBoost and LightGBM, 1000 iterations of Bayesian HPO were performed, as by Schäfl et al. [35]. As CatBoost requires significantly more training time, only 100 iterations of Bayesian HPO were performed. We also used a logistic regression without HPO.

For each test fold, the AUROC and feature importance (FI) were determined. FI was calculated for each feature individually and for clusters of features if they exhibited high collinearity, as it affects the importance value. The identification of feature clusters was based on [https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html]. The FI was calculated using the scikit-learn package and its permutation FI implementation. The method was modified, to simultaneously permute clusters of features if they exhibited high collinearity.

Missing values of numerical variables were imputed with the mean value of the respective training folds. All numerical variables were z-standardized. The mean and standard deviation were obtained from the respective training folds before imputation.

2.4. Frailty Model Validation

To ensure that our best-trained model captured the frailty concept, we followed Rockwood [36] with respect to construct validity and criterion validity. For construct validity, we used HFRS. Criterion validity was assessed by predicting adverse outcomes like falls during stay, delirium during stay and mortality during stay. For model validation, the GBDT model with the best validation score was used. For each test sample, a prediction was performed. These prediction values were then correlated with HFRS. Additionally, the prediction values were z-standardized and then used to predict adverse outcomes with the help of a logistic regression. The model was validated on each test fold.
3. Results

3.1. Frailty Prediction

XGBoost had the best model performance in terms of validation score (see Table 1). It achieved an average AUROC of 0.696 across all test folds. The average AUROC per center was 0.696 (XGBoost). The AUROC weighted by the number of samples per facility resulted in 0.675 (XGBoost).

Table 1. Comparison of individual model performances with respect to the prediction of frailty. The values are the means and 95% confidence intervals (in brackets) of the AUROCs. One AUROC is generated per method for each run of the 5-fold CV.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test (Mean, CI)</th>
<th>Validation (Mean, CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>XGBoost</td>
<td>0.696 (0.687,0.705)</td>
<td>0.701 (0.699,0.703)</td>
</tr>
<tr>
<td>LightGBM</td>
<td>0.695 (0.687,0.703)</td>
<td>0.700 (0.697,0.703)</td>
</tr>
<tr>
<td>CatBoost</td>
<td>0.689 (0.681,0.697)</td>
<td>0.699 (0.696,0.702)</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.666 (0.655,0.677)</td>
<td>0.667 (0.663,0.671)</td>
</tr>
</tbody>
</table>

3.2. Important Features

The most important model features and feature clusters are shown in Figure 1. The bars indicate how much the AUROC decreases if the values of a feature or a feature cluster are randomly shuffled. The relevant feature clusters are listed in Table 2.

Figure 1. Mean feature importance across all test folds. The black lines represent the standard deviation.

Table 2. Important clusters with respect to frailty prediction. Abbreviations: CRP: C-reactive protein; LDH: lactate dehydrogenase; GPT: glutamic pyruvic transaminase; GOT: glutamic oxaloacetic transaminase; MCHC: mean corpuscular hemoglobin concentration; rel.: relative; abs.: absolute.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>age, potassium, urea, creatinine, uric acid</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>eosinophilic granulocytes rel., basophilic granulocytes rel., eosinophilic granulocytes abs., basophilic granulocytes abs.</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>sodium, chloride</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>hematocrit, hemoglobin, erythrocytes, CRP, calcium</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>LDH, GPT, GOT</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>neutrophilic granulocytes rel., lymphocytes rel., lymphocytes abs.</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>leukocytes, monocytes rel., neutrophilic granulocytes abs., monocytes abs.</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>thrombocytes, MCHC, bilirubin</td>
</tr>
</tbody>
</table>
The following laboratory values are not listed in Table 2 because of their low feature importance: mean corpuscular volume, mean corpuscular hemoglobin, gamma-glutamyltransferase (GGT), glucose, prothrombin time, alkaline phosphatase, activated partial thromboplastin time.

Cluster 1 is the most significant for model performance. A large part can be explained by age. However, the rest of the cluster is important to increase model performance. Urea and creatinine might still play an important role. The granulocytes also contribute to model performance, as does chloride and calcium.

### 3.3. Frailty Model Validation

The XGBoost model predictions achieved a mean correlation of $r=0.146$ (min $r=0.129$, max $r=0.166$) with the HFRS across all test folds (all $p<0.001$).

Predicting mortality during stay resulted in a mean standardized regression coefficient of $\beta_1=0.757$ (min $\beta_1=0.593$, max $\beta_1=0.928$; all $p<0.0136$). The prediction achieved an AUROC of $0.756\pm0.028$. Prediction of delirium resulted in a mean $\beta_1=0.311$ (min $\beta_1=0.035$, max $\beta_1=0.545$; at least one $p>0.8$). Prediction of falls resulted in a mean $\beta_1=0.119$ (min $\beta_1=-0.206$, max $\beta_1=0.289$; at least one $p>0.4$).

### 4. Discussion

#### 4.1. Interpretation

The XGBoost model achieved an AUROC of 0.696 for frailty prediction. This performance might be too low for clinical use. The most important features were age, urea, creatinine, granulocytes, chloride and calcium. It is not surprising that age predicts frailty, as frailty prevalence increases with age [4]. C-reactive protein might be important for the model (see Figure 1) and is associated with frailty [37, 38].

The XGBoost model predictions were correlated with HFRS and can be used to predict mortality during stay. The model predictions hold no significant information to predict falls and deliriums during stay. Therefore, the criterion validity is unclear.

In this study, laboratory values such as albumin and 25-hydroxyvitamin D were not taken into account, as they are not routinely collected. However, these two laboratory values have been associated with frailty [37, 38].

It must be noted that patients might already have received medical treatment when they were transferred to geriatrics from another ward or hospital. This might influence laboratory values. Therefore, considering the history of laboratory values could be important. The question arises whether infrequently collected laboratory values or perhaps laboratory value history could improve model performance.

#### 4.2. Limitation

Our developed model was based on hospitalized older individuals already admitted to geriatric wards. Whether this model is also valid for hospitalized older individuals outside of geriatric wards still has to be tested.

Predicting adverse outcomes for frail individuals in geriatric wards is difficult, as these individuals receive adequate care and adverse outcomes are less likely to occur. In
addition, the stay in geriatric wards is often too short for adverse outcomes to occur. Consequently, the validation of the frailty model with the help of the chosen adverse outcomes yields no strong evidence.

Age is a strong predictor of frailty. The question arises, how much information with regard to frailty is contained in the laboratory values alone.

After model development, we tested the independence between the availability of laboratory values and frailty using a $\chi^2$-Test. Eight laboratory values had a $p<0.05$ (five absolute values of the different leukocyte types, GGT, LDH, GOT). Subsequently we performed the experiment without these eight variables and obtained an AUROC of 0.690. Thus, exclusion of the variables makes little difference. The significance probably is caused by different laboratory routines in the individual facilities in combination with a different distribution of frailty between the facilities.

4.3. Conclusion

We trained a ML model to identify frailty in older individuals in the hospital with the help of laboratory values. The model is quick, requires no additional effort by healthcare personnel and is applicable in practice on a large scale. The performance is acceptable, but insufficient for clinical use. Additional laboratory values or laboratory value history could improve the performance. We identified urea, creatinine, granulocytes, chloride and calcium as important laboratory features to predict frailty.

The validation of the model as a frailty assessment is not clear. The model should be validated on hospitalized older individuals outside the geriatric ward. These individuals might not have received adequate therapy if they are frail. The model should be tested to see how well it can predict adverse outcomes for these individuals.

4.4. Author Statement

Research funding: FFG, IdRis project #888068. Ethical approval: Ethics committee of the Medical University of Graz, 34-397 ex 21/22, Univ. Prof. DI Dr. Josef Haas.

References


