Heva
Une filiale de DOCAPOSTE

Nicolas Pagès[1], Arnaud Panes[1], Stanislas Perrier[2], Jules Tavi[2], Anne-Lise Vataire[2], Margaux Gouysse[1], Aurélie Schmidt[1]

1. Heva, 186 Avenue Thiers, Lyon, France; 2. Sanofi, 82 avenue Raspail, Gentilly, France

ISPOR Europe 2024
17-20 November 2024 - Barcelona, Spain

EE474

Valuation and exploitation of sick leaves identified in the French SNDS database

INTRODUCTION

The analysis of sick leaves (SL), particularly within the French National Health Data System (SNDS), prompts several questions about effectively describing and leveraging these direct costs, which pose a significant economic burden in many diseases.

A study was conducted on patients undergoing triple therapy (marker of severity) for chronic obstructive pulmonary disease (COPD), aiming to describe and analyze those with sick leaves.

METHODOLOGY

A retrospective observational study was conducted using the SNDS. Adult patients (≥ 40 years) treated with triple therapy for COPD between January 1, 2015, and December 31, 2015 were included.

A one-year follow-back and a 6-year follow-up period was used to identify sick leaves through daily allowances paid by the French National Health Insurance (NHI).

Non-retired patients aged 65 year-old or younger with at least one sick leave were then analyzed using descriptive analyses, a graphical representation method based on patient clustering (TAK® methodology) and a machine learning approach to predict and explain sick leaves.

CONCLUSION

This study underscores the significant impact of sick leaves among COPD patients.

The findings reveal that a notable proportion of these patients require substantial time off work, with over 20% experiencing prolonged sick leaves exceeding one year with an increased financial burden on the NHI system.

Additionally, the use of advanced methodologies, such as TAK® for patient clustering or machine learning approach, offers valuable insights into the patterns and predictors of sick leaves.

Abbreviations

AUC: area under the curve; **LS:** sick leaves; **NHI:** French National Health Insurance; **NPV:** negative predictive value; **PPV:** positive predictive value; **SHAP:** SHapley Additive exPlanations; **SNDS:** French National Health Data System

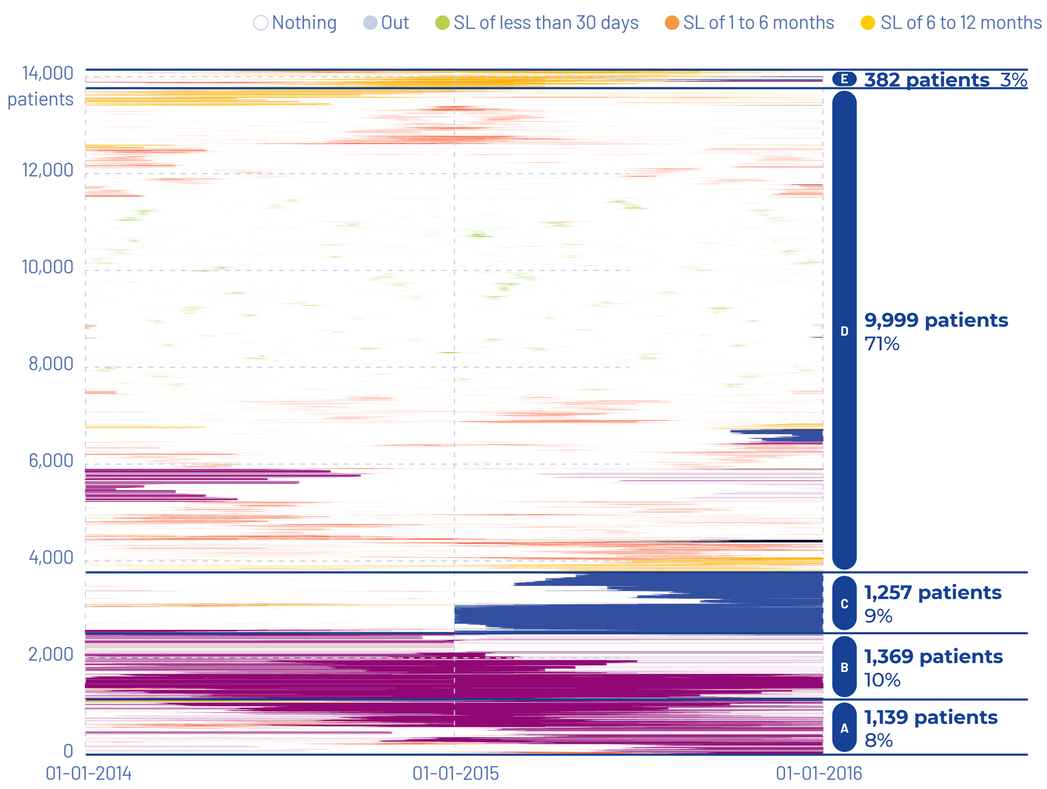
Data Sources

SNIIRAM study registered with the HDH on 06/30/2022 and authorised by the CNIL on 05/09/2022. (DR-2022-194 (request 922190)) - CNAM agreement signed on 07/12/2023.

RESULTS

186,963 COPD patients treated by triple therapy were included. Of these, 63% were men, and 54,459 were non-retired patients aged 65 year-old or younger. Over the study period, 64,270 SL were recorded in our population.

TAK®



Methodology

All patient with a SL between 2014 and 2015 were included in this analysis (N=14,352).

SL were then categorized by their duration (< 30 days, 1 to 6 months, 6 to 12 months and more than one year).

Results

Among our COPD patients, more than 20% had a SL of more than year. Most patients had SL of less than 6 months.

This chart shows that this clustering method could be useful for representing sick leaves. In this analysis, we can observe a significant number of patients with prolonged sick leave durations, which could be related to the burden of the pathology.

How to read the TAK®

The TAK® graph represents follow-up of each patient of the cohort in a linear graph. Each patient starts on the 1st of January 2014 on the vertical axe at the left of the graph. Then it moves horizontally to the right, over the course of the 2-year follow-up (12-31-2015).

Machine learning

1 Data preprocessing

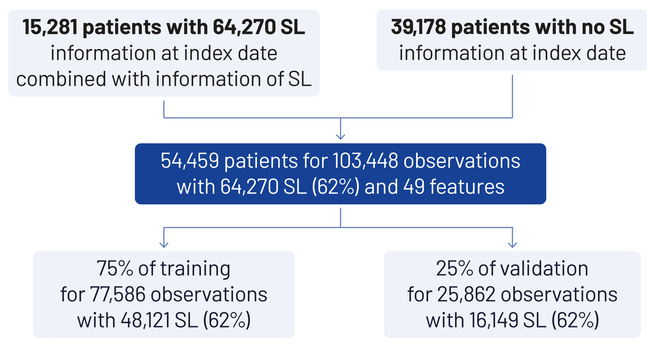
Methodology

The database was created by compiling data on sick leaves, with each row representing one instance of sick leave or information at index date. Key candidate variables identified for analysis included the patient's age at the time of the SL, the number of exacerbations in the year preceding the SL event, the Charlson comorbidity index score, and any relevant comorbidities.

Detailed methodology

To ensure robust model training and validation, the dataset was split into training and validation sets using a 75/25 ratio, while maintaining the proportion of SL events across both sets. Rigorous checks were performed on the data split to prevent any potential biases.

Results



2 Algorithms processing

Methodology

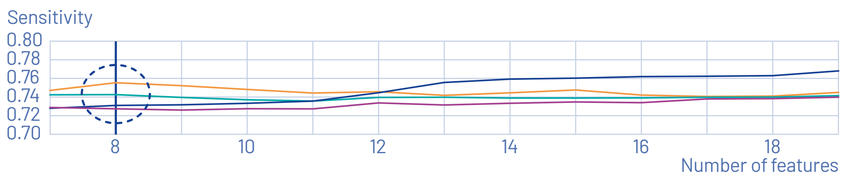
To create a model suitable for routine use, the number of selected variables was optimized between 7 and 19. Four models were compared and the performance of the models was evaluated using multiple metrics.

Detailed methodology

Several univariate feature selection methods were tested, with the ANOVA f-value score ultimately chosen for its effectiveness. This method select features using the degree of linear dependency between the feature and the outcome.

Results

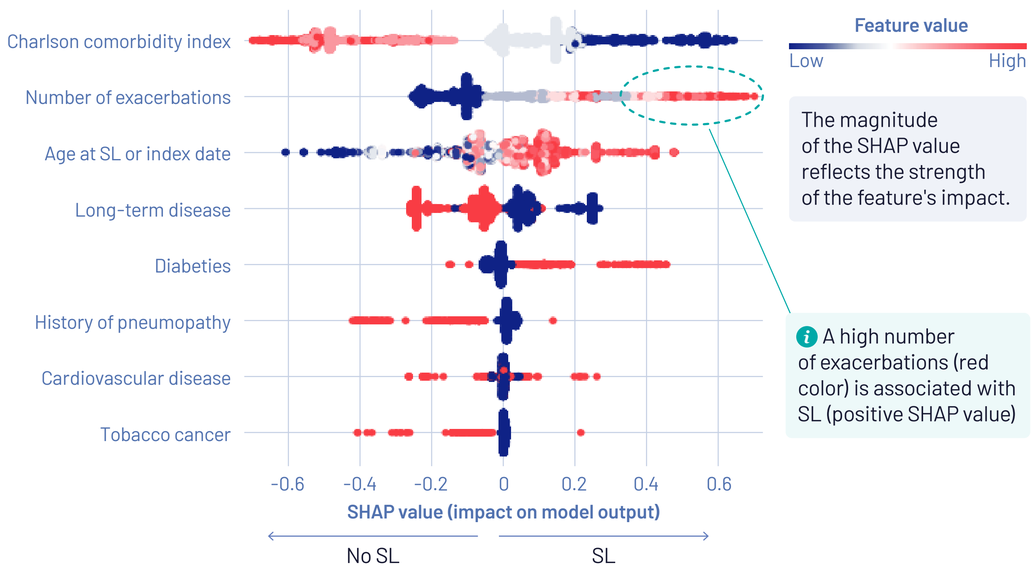
The Random Forest model with 8 selected variables was chosen as the optimal model, as it provided the best performance without a significant gain in sensitivity when additional variables were included.



On the validation dataset, the Random Forest model successfully identified **83%** of sick leave events, with a positive predictive value (PPV) of **75%**, confirming its efficacy in this context.

	Sensitivity	Specificity	VP	VPN	AUC	F1-score
Random forest	0.756	0.734	0.823	0.648	0.835	0.788
Logistic regression	0.743	0.719	0.812	0.631	0.808	0.776
Gradient boosting	0.728	0.768	0.837	0.633	0.838	0.778
SVM	0.731	0.755	0.830	0.632	0.768	0.777

3 Algorithm behavior



A positive SHAP value indicates that a feature increases the predicted outcome (more SL in this case), while a negative SHAP value shows that the feature decreases it (less SL).

Methodology

To gain insights into the model's behavior, SHAP (SHapley Additive exPlanations) values were calculated for a sample of 1,000 sick leave events.

SHAP values are a game-theory-based method used to explain the predictions of machine learning models. In practice, SHAP values are used to quantify the effect of each feature on an individual prediction.

Results

The analysis revealed that a high number of exacerbations in the preceding year or being diabetic emerged as predictors of sick leave, indicating their critical role in forecasting these events.

Conversely, certain factors, such as an important Charlson comorbidity index or having a long-term disease seemed to be less associated with SL suggesting that those patients were probably no longer working.

Discussion

This model can be a pertinent tool to predict and explain sick leaves with patients' characteristics and medical history. However, it does not account for the potential evolution of comorbidities and the Charlson score over time, which could affect the accuracy of predictions. Another limitation is the inability to identify patients who are not actively working, which could impact the model's generalizability.