Poster #MSR118

Identifying potential long COVID patients using machine learning: A German claims data analysis

Scarlette Pacis[1], Anna Bolzani[1], Ulf Maywald[2], Thomas Wilke[3]
[1]Cytel, Inc., Berlin, Germany; [2]AOK PLUS, Dresden, Germany; [3]Institut für Pharmakoökonomie und Arzneimittellogistik e.V. (IPAM), Wismar, Germany

Machine learning can be used to determine important features for identifying potential long COVID patients, including new diagnoses and healthcare resource utilization following initial COVID-19 hospitalization.

Background

- Long COVID symptoms include a wide range of new health problems following COVID-19 infection that can last three or more months after the first onset of post-COVID symptoms.[1]
- Commonly reported symptoms include fatigue, dyspnea, cognitive and mental impairments, smell and taste dysfunctions, chest and joint pains, cough, headache, and other gastrointestinal and cardiovascular disorders.[2]
- Many patients suffering from long COVID may have experienced delayed diagnosis and received untimely treatment, especially during the beginning of the pandemic.

Objective

- This study uses machine learning to determine important features for identifying potential long COVID patients as a proxy for long COVID diagnosis.

Methods

Data Source and Patient Selection

- Data from AOK PLUS, a German sickness fund covering 3.5 million patients in Saxony and Thuringia were used.
- All adult patients with ≥1 inpatient documentation of confirmed COVID-19 (ICD-10-GM: U07.1) between 01/04/2020-31/03/2022 (index date = first COVID-19 diagnosis), alive at 31/03/2022, and with ≥90 days continuous insurance after index were included.
- The outcome of interest was ≥1 long COVID diagnosis (inpatient/outpatient; U09.9!) during follow-up (45-365 days after index, or to long-COVID diagnosis).

Machine Learning Model

- An XGBoost model (70/30 training/testing) was developed with 207 initial features including characteristics at index (age, sex, intubation, comorbidities, Charlson-comorbidity score [CCI]), any new diagnoses and medications during 30-365 days after index that did not occur in 30-365 days before index, and healthcare utilization (number of outpatient visits/hospitalization days in follow-up) (Figure 1, Table 1).
- Shapley values were used for feature interpretability and the final model included the top 25 most important features.
- SMOTE (Synthetic Minority Oversampling Technique) was used to improve model performance due to the moderate class imbalance.
- Model performance was assessed using AUROC, sensitivity, specificity, precision, recall, and F1 score.

Figure 1. Temporal windows for model inclusion

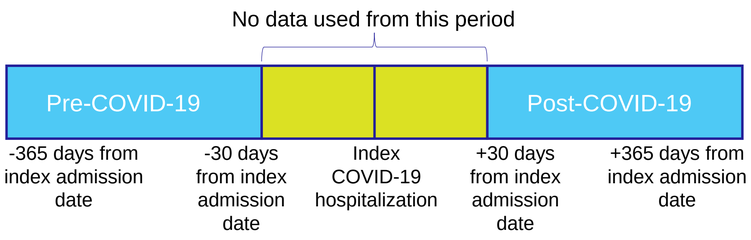


Table 1. Initial features considered for model inclusion

Feature	Description
Baseline characteristics	Age, sex, CCI, ICU or intubation during index hopsitalization
Pre-COVID comorbidities	Binary variables to flag whether each patient had at least 1 inpatient or 2 outpatient diagnosis in pre-COVID window for: diabetes [E10-E14], CKD [N18], CHF [I50], COPD [J44].
Healthcare utilization ratios	Ratio of healthcare utilization in post-COVID-19 window to number of days in patient's post-COVID-19 window[1]
Post-COVID diagnoses	Binary variables of diagnoses (outpatient or inpatient [any position]) that newly occurred in the post-COVID-19 period and during the pre-COVID-19 period. Identified using ICD-10-GM codes up to 3 characters.[2]
Post-COVID prescriptions	Binary variables of prescriptions that newly occurred in the post-COVID-19 period and not during pre-COVID-19 period. Identified using ATC codes up to 4 characters.

[1]Outpatient utilization ratio = Number of outpatient visits based on in post-COVID-19 window / Number days in post-COVID-19 window. Number of outpatient visits is counted as the number of different days at which an EBM code was billed. Inpatient utilization ratio = sum of lengths of stay of all inpatient visits in post-COVID-19 window / Number of days in post-COVID-19 window

[3]Include only diagnoses/prescriptions associated with at least 1% of patients

Results

- 28,419 patients were included, of which 6,512 (22.9%) patients had long COVID (Figure 2).
- The proportions of patients aged 45-64 and 65-74 years at index were larger for long COVID patients compared to non-long COVID patients (Table 2).
- Long COVID patients are more likely to have received treatment in the intensive care unit (ICU) and intubation during the index hospitalization for COVID-19.
- Pre-COVID comorbidities of diabetes, chronic kidney disease (CKD), and chronic heart failure (CHF) were more common among non-long COVID patients, while chronic pulmonary disease (CPD) was more common long COVID patients.
- AUROC, sensitivity/specificity, and F1 score were 0.80, 0.67/0.93, and 0.70, respectively (Figure 3).
- Outpatient and inpatient healthcare resource utilization (HCRU) had the largest impact on the model (Figure 4).
- Patients with new symptoms of viral pneumonia, breathing disorders, respiratory failure, and malaise and fatigue were more likely to be identified by the model as having long COVID.
- Patients with new prescriptions of analsegics/antipyretics, peptic ulcer drugs, anti-inflammatory/anti-rheumatic, antipsychotics, and high-ceiling diuretics were less likely to be identified by the model as having long COVID.

Figure 2. Cohort Selection

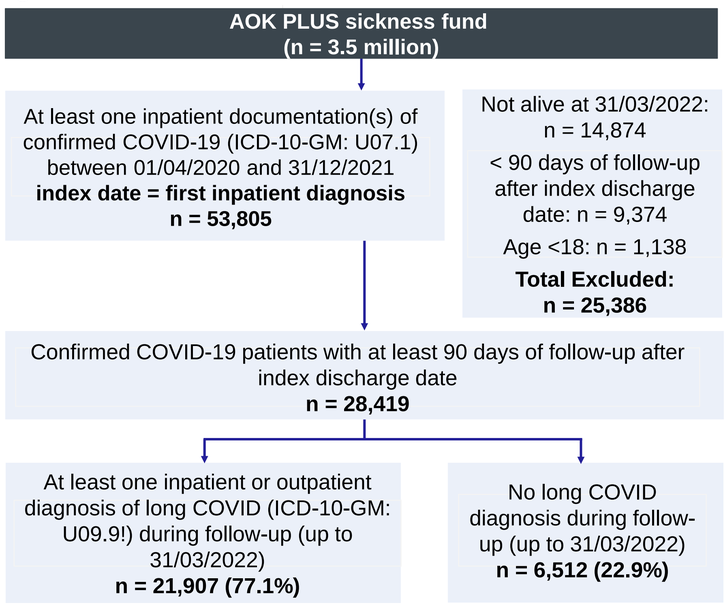


Figure 3. Confusion matrix and evaluation metrics

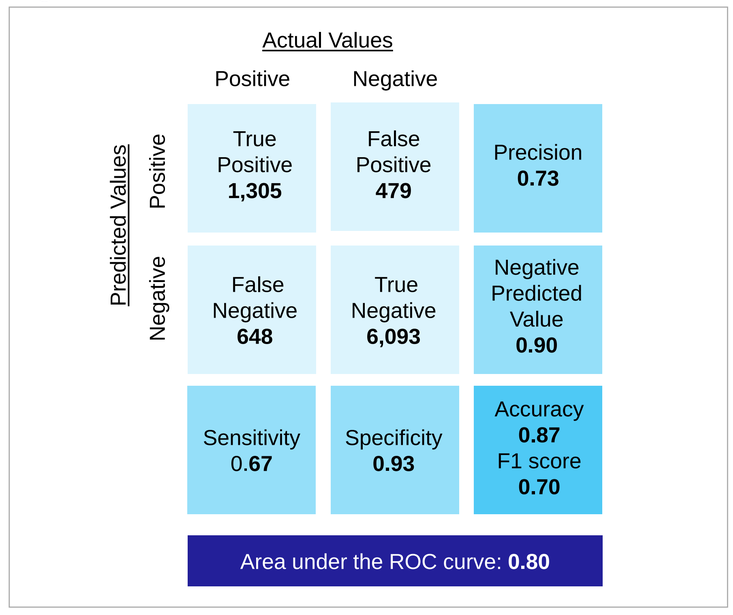
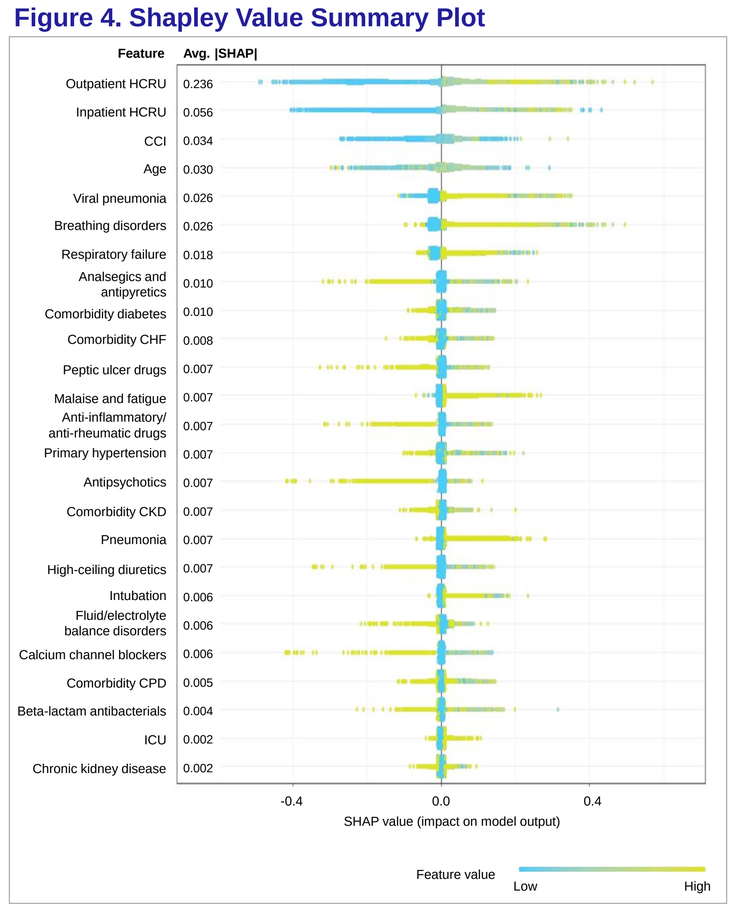


Table 2. Pre- and Post-COVID Characteristics

	All COVID-19 patients (n=28,419)	Long COVID diagnosis (n=6,512)	No Long COVID diagnosis (n = 21,907)	p-value
Baseline characteristics				
Follow-up in months, mean (SD)	11.4 (5.3)	12.0 (5.4)	11.2 (4.8)	<0.001
Age, mean (SD) at index date	66.5 (17.7)	65.1 (18.9)	67.0 (18.3)	<0.001
<45 years, n (%)	3,913 (13.8)	692 (10.6)	3,221 (14.7)	<0.001
45 to 64, n (%)	7,616 (26.8)	2,416 (37.1)	5,200 (23.7)	-
65 to 74, n (%)	5,548 (19.5)	1,383 (21.2)	4,165 (19.0)	-
75 to 84, n (%)	7,107 (25.0)	1,434 (22.0)	5,673 (25.9)	-
85+	4,235 (14.9)	587 (9.0)	3,648 (16.7)	-
Female, n (%)	15,349 (54.0)	3,22 (51.0)	12,02 (54.9)	<0.001
CCI	3.6 (3.2)	3.7 (3.2)	3.6 (3.2)	0.099
ICU	1,302 (4.6)	654 (10.0)	648 (3.0)	<0.001
Intubation	987 (3.5)	586 (9.0)	401 (1.8)	<0.001
Pre-COVID-19 comorbidities, n (%)				
Diabetes [E10-E14]	9,725 (34.2)	2,146 (33.0)	7,579 (34.6)	0.014
CKD [N18]	6,237 (22.0)	1,327 (20.4)	4,910 (22.4)	<0.001
CHF [I50]	5,821 (20.5)	1,198 (18.4)	4,623 (21.1)	<0.001
CPD [J44]	2,907 (10.2)	734 (11.3)	2,173 (9.9)	0.002
Post-COVID-19 characteristics				
HCRU utilization ratios (mean percentage, SD)				
Outpatient	2.1 (2.1)	2.2 (1.8)	1.5 (2.8)	<0.001
Inpatient	6.2 (88.6)	20.5 (184.1)	1.9 (5.0)	<0.001
Post-COVID diagnoses – Top 5 features [ICD-10-GM] (n, %)				
Viral Pneumonia [J12]	3,157 (11.1)	1,190 (18.3)	1,967 (9.0)	<0.001
Breathing Disorders [R06]	2,590 (9.1)	945 (14.5)	1,645 (7.5)	<0.001
Respiratory Failure [J96]	3,477 (12.2)	1,258 (19.3)	2,219 (10.1)	<0.001
Malaise and Fatigue [R53]	1,608 (5.7)	453 (7.0)	1,155 (5.3)	<0.001
Primary Hypertension [I10]	5,070	1,055 (16.2)	4,015 (18.3)	<0.001
Post-COVID prescriptions – Top 5 features [ATC] (n, %)				
Analsegics and Antipyretics [N02B]	3,516 (12.4)	362 (5.6)	3,154 (14.4)	<0.001
Peptic ulcer drugs [A02B]	3,070 (10.8)	408 (6.3)	2,662 (12.2)	<0.001
Anti-inflammatory/ anti-rheumatic [M01A]	1,964 (6.9)	147 (2.26)	1,817 (8.3)	<0.001
Antipsychotics [N05A]	1,470 (5.2)	107 (1.6)	1,363 (6.2)	<0.001
High-ceiling diuretics [C03C]	2,104 (7.4)	235 (3.6)	1,869 (8.53)	<0.001

Figure 4. Shapley Value Summary Plot



Conclusions

- Patients with more outpatient and inpatient HCRU after initial COVID-19 hospitalization were more likely to be identified by the model as having long COVID.
- New symptoms of viral pneumonia, breathing disorders, respiratory failure, and malaise and fatigue were most associated with diagnosis of long COVID.

Abbreviations

ATC: Anatomic therapeutic chemical; CCI: Charlson Comorbidity Index; CKD: Chronic kidney disease; CHF: Chronic Heart Failure; COPD: Chronic pulmonary disease; COVID: Severe acute respiratory syndrome coronavirus; EBM: Uniform Valuation Standard *"Einheitlicher Bewertungsmaßstab";* HCRU: Healthcare resource use; ICD-10-GM: International classification of diseases Germany; ICU: Intensive care unit; ROC: Receiver Operating Characteristics; SHAP: Shapley value; SMOTE: Synthetic Minority Oversampling Technique

References

1. Yong SJ. Long COVID or post-COVID-19 syndrome: putative pathophysiology, risk factors, and treatments. Infect Dis (Lond). 2021 Oct;53(10):737-754. doi: 10.1080/ 23744235.2021.1924397.
2. van Kessel SAM, Olde Hartman TC, Lucassen PLBJ, van Jaarsveld CHM. Post-acute and long-COVID-19 symptoms in patients with mild diseases: a systematic review. Fam Pract. 2022 Jan 19;39(1):159-167. doi: 10.1093/fampra/cmab076.

Disclosures

No conflicts of interest are reported. No funding was received for this work.

Presented at ISPOR EU | Copenhagen, Denmark | November 12–15, 2023

Contact information: www.cytel.com