Statistical and Predictive analysis to Identify Risk Factors and Effects of Post COVID-19 Syndrome

Milad Leyli-abadi*, Jean-Patrick Brunet+

Sonja van Ockenburg+, Axel Tahmasebimoradi*

IRT SystemX, Palaiseau, France *

University Medical Center Groningen, Netherlands +

SystemX INSTITUT DE RECHERCHE TECHNOLOGIQUE

University Medical Center Groningen

IARIA

Outline

Introduction

Context and motivation

Data and statistical analysis

Predictive analysis of long COVID intensity

Conclusions and perspectives

Context & Motivation

In **May 2023**, the WHO declared the end of the **COVID-19 global Public Health Emergency**, marking a new endemic phase

However, **COVID-19 continues to have lasting effects**, notably **Post-COVID-19 Condition (PCC)** — persistent physical and cognitive symptoms following recovery

Recent global estimates suggest a **6–10% prevalence**, down from initial 10–20% WHO estimates

Identifying **populations at risk of PCC** is essential for **early referral** and optimized care pathways

Research Challenges & Data Opportunity

PCC characterization remains uncertain, with key challenges:

- Incomplete, retrospective data collected during the evolving crisis
- Focus on hospitalized patients, despite most PCC cases occurring in **non-hospitalized individuals**
- Limited information on **pre-existing conditions**, complicating symptom attribution

Lifelines Cohort offers a unique opportunity:

- **167,729 participants** from Northern Netherlands across three generations
- COVID-19 branch (Apr 2020 – Nov 2022):
  - 31 questionnaires, weekly to bi-monthly
  - **76,503 respondents**, average 13.5 questionnaires each
- Rich longitudinal data enable **analysis of pre-infection factors** and PCC trajectories

Aim & Contributions

Research Question:

- ***Can pre-infection parameters predict the severity of Post-COVID-19 Condition?***

Approach:

- Introduce **Post-COVID-19 Symptom Intensity (PCSI)** as a **continuous measure** of PCC severity
- Apply **machine learning models** to Lifelines data to predict PCSI and identify risk factors

Key Contributions:

- Statistical identification of influential factors associated with PCC
- Predictive modeling of PCSI using data-driven techniques
- Interpretation of variable impact for medical decision-making
- Development of an **open-source Python package** for reproducible ML analysis

Dataset Overview

Two main variable types:

- **Static variables:** Fixed individual attributes (e.g., age, sex, variant, income, smoking, chronic diseases, vaccination status, time between vaccination and infection)
- **Dynamic variables:** Symptom presence and intensity over time (before, during, and after infection)
  - Includes headache, cough, fever, breathing difficulties, muscle pain, loss of smell/taste, etc.

Data Challenges:

- Substantial **missing and inconsistent data** (common in questionnaire-based studies)
- **Evolving questionnaire design** during the epidemic → variable phrasing and coverage
- Required **extensive standardization** to build a uniform analytical dataset

Defining Post-COVID-19 Symptom Intensity (PCSI)

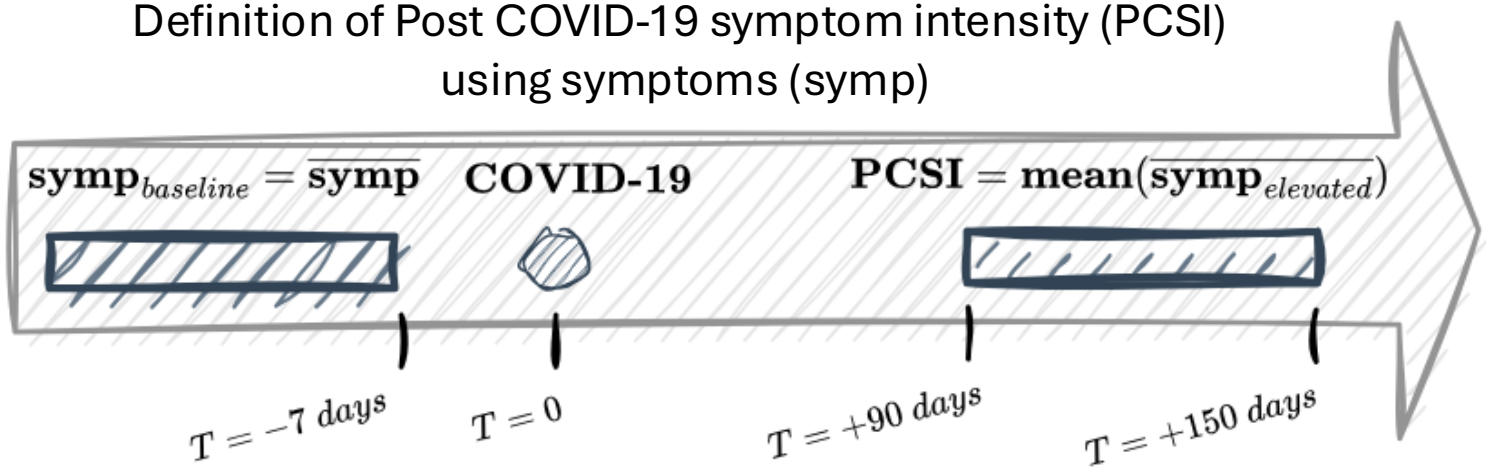
Definition of Post-Covid Condition based on the WHO one

Symptoms lasting more than 3 months post infection lasting for at least 2 months with no other explanation.

Use of 10 symptoms checked to be related to PCC in a previous study

Extension to a continuous variable by deriving the Post Covid-19 symptom intensity score (PCSI)

Definition of Post COVID-19 symptom intensity (PCSI) using symptoms (symp)



Preprocessing Workflow & Sample Selection

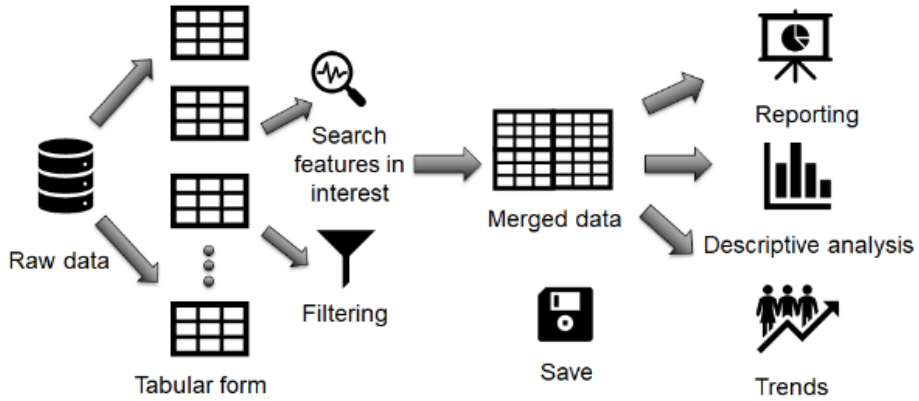
Data Integration:

- Raw questionnaire data organized by participant and date
- Merged across tables after cleaning and variable harmonization
- Feature extraction based on **pre-infection period** (steady-state hypothesis)

Final Study Sample:

- **4,657 participants** included after filtering for sufficient data
- Comparable baseline characteristics between included and excluded groups
- **Sex distribution:**
  - Women = 73% of PCC cases vs. 64% in base dataset → higher PCC risk in women
  - Lower PCSI levels show a reduced female proportion

Data preparation strategy



Statistical Tests

Tests Used:

- **Chi-square test** → checks independence between categorical variables (e.g., vaccination & PCSI
- **Cramer's V** → measures **strength** of association (0 = weak, 1 = strong)

Vaccination & PCSI:

- Majority of **fully vaccinated** participants had **low PCSI** (88% scored 1–2)
- **Chi-square:** significant relationship (**p < 0.05**)
- **Cramer's V = 0.072:** weak association strength

Insight:

- Vaccination status is significantly associated with lower PCC intensity, but the strength of this association is relatively small in practical terms

$$\begin{cases} H_0: Independence\ between\ variables \\ H_1: Dependence\ between\ variables \end{cases}$$

VACCINE	PC_INTENSITY 1	2	3	4	5	**Total**
complete vaccin	2514 79.8 % 71 %	276 8.8 % 54.9 %	225 7.1 % 59.7 %	108 3.4 % 54.5 %	26 0.8 % 70.3 %	3149 100 % 67.6 %
no	1028 68.2 % 29 %	227 15.1 % 45.1 %	152 10.1 % 40.3 %	90 6 % 45.5 %	11 0.7 % 29.7 %	1508 100 % 32.4 %
Total	3542 76.1 % 100 %	503 10.8 % 100 %	377 8.1 % 100 %	198 4.3 % 100 %	37 0.8 % 100 %	4657 100 % 100 %

χ^2=81.995 · df=4 · Cramer's V=0.133 · p=0.000

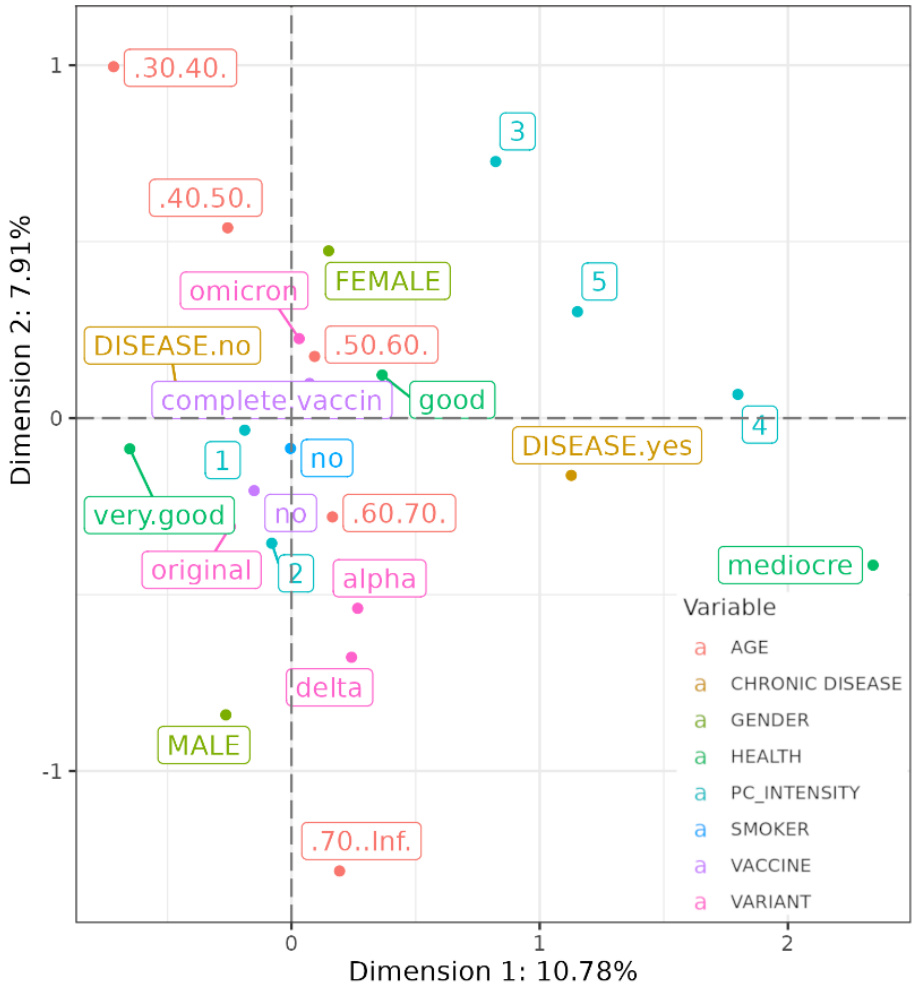
Multiple Correspondence Analysis (MCA)

Method

- **MCA** was applied to explore **simultaneous relationships** between multiple **categorical variables**

Key patterns

- **High PCSI (5)** clusters with **chronic diseases**, **poor health**, and **female sex**, indicating higher vulnerability in these groups
- **SARS-CoV-2 variant** shows **weak association** with PCC, suggesting a lower impact on long-term symptom intensity
- Participants with **better general health** are positioned closer to **low PCSI levels**, indicating a protective effect

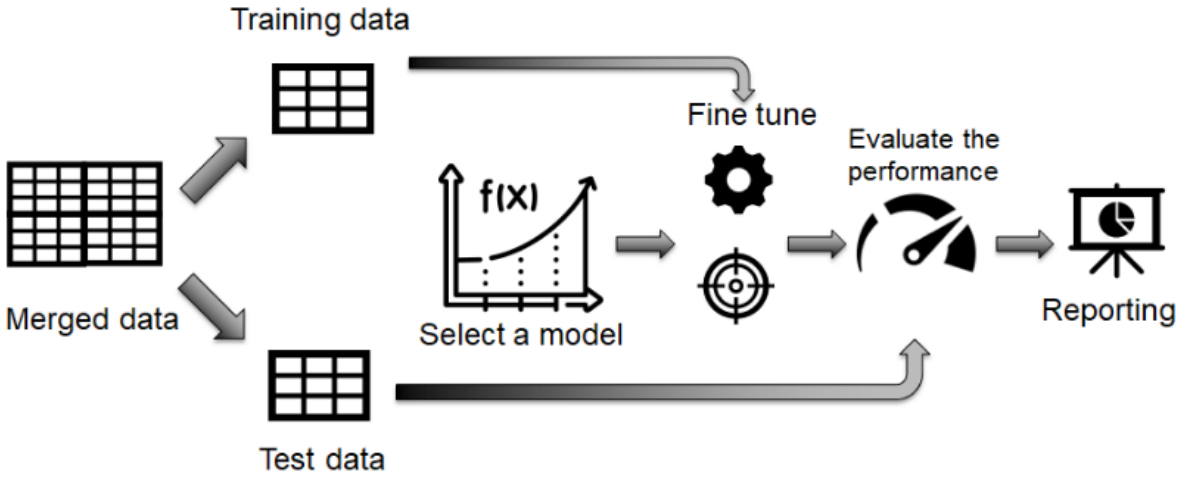


Predictive analysis

The model $f: \boldsymbol{X} \rightarrow y$

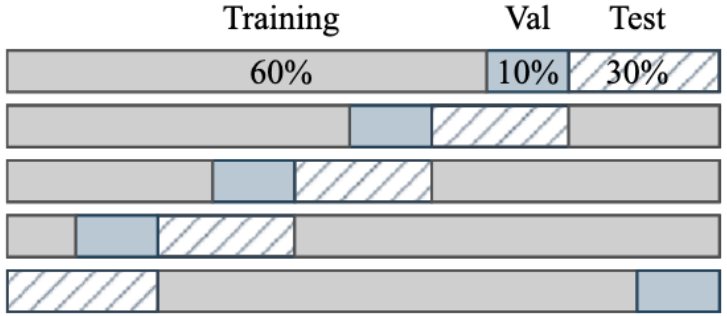
- With $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{p}}$ representing the explanatory variables (features) of dimension p
- And y representing the target variable corresponding to the long covid intensity with $y \in [1; 5]$

Evaluation pipeline



- To be robust to data variation during the training and evaluation , the cross-validation strategy is adopted
- The final results are reported using mean and standard deviation over multiple folds

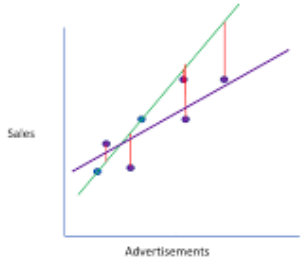
Cross validation



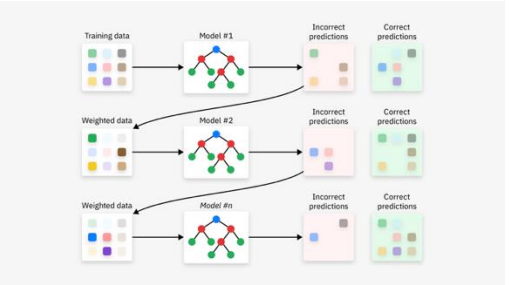
Methods and evaluation criteria

Evaluated methods

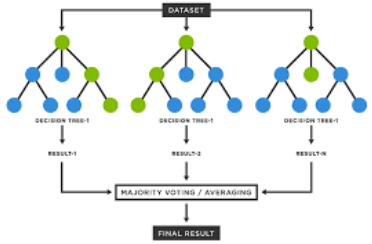
- Linear Ridge Regression (LR)



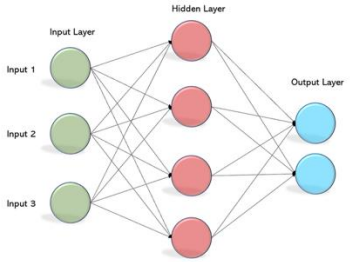
- Gradient Boosting (GB)



- Random Forest (RF)



- Multi-layer Perceptron (MLP)



Evaluation criteria

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Mean Absolute Percentage Error (MAPE)
- Pearson Correlation

Results

- **Best performance achieved with "All" features** across all methods
- **Symptom-only features** yield similar results for LR, RF, GB; MLP benefits more from full feature sets
- **MAE ≈ 0.60** for all models → acceptable error on a 1–5 PCSI scale
- **MLP** shows **lowest MAPE (0.19)**, minimizing relative errors
- **RF and GB** yield the **highest Pearson correlations**, capturing complex linear interactions

Differences reflect **model strengths**:

- RF/GB: strong linear patterns → higher correlations
- MLP: non-linear modeling → better error minimization

		Evaluation criteria			
Methods	Features	MAE	MSE	MAPE	Pearson
LR	All	**.61 ± .01**	**.68 ± .02**	**.29 ± .01**	**(.56, 6e-70)**
	Static	.71 ± .02	.91 ±.05	.35 ± .01	(.28, 2e-16)
	Symptoms	.62 ± .02	.70 ± .04	.30 ± .01	(.57, 2e-69)
	Vaccination	.81 ± .02	.99 ± .05	.41 ± .01	NaN
RF	All	**.60 ± .01**	**.67 ± .02**	**.28 ± .01**	**(.58, 7e-73)**
	Static	.72 ± .02	.93 ± .05	.35 ± .01	(.26, 1e-15)
	Symptoms	**.60 ± .01**	.66 ± .03	**.28 ± .01**	(.57, 5e-72)
	Vaccination	.79 ± .02	.99 ± .06	.39 ± .01	(.04, 1e-1))
GB	All	**.61 ± .01**	**.66 ± .01**	**.28 ± .01**	**(.57, 4e-74)**
	Static	.72 ± .02	.90 ± .05	.35 ± .01	(.29, 7e-17)
	Symptoms	**.61 ± .01**	.68 ± .02	.28 ± .01	(.55, 8e-82)
	Vaccination	.81 ± .02	.99 ± .06	.41 ± .01	(.05, 6e-1)
MLP	All	**.45 ± .05**	**.90 ± .12**	**.19 ± .03**	**(.25, 3e-18)**
	Static	.87 ± .18	1.4 ± .78	.43 ± .07	(.21, 4e-9)
	Symptoms	.76 ± .11	.98 ± .38	.34 ± .05	(.43, 5e-33)
	Vaccination	.80 ± .03	1.03 ± .05	.41 ± .03	(.04, 2e-1)

COMPARISON BETWEEN VARIOUS INTRODUCED MODELS AND FEATURES COMBINATION FOR PREDICTION OF PCSI

Interpretation

Linear Regression

Top 9 features identified via **Linear Regression**

- Coefficients show **direction & strength** of impact on **PCSI**
- **Positive predictors** (↑ PCC risk):
  - **Loss of smell, headache, muscle pain**
- **Negative predictors** (↓ PCC risk):
  - **Fever, pain when breathing**

Insights

- Acute symptoms differ in **predictive value** — some signal higher long-term risk, others appear protective

Variable	Coef	Variable	Coef
Loss of sense of smell/taste	0.32	Pain when breathing	-0.58
Headache	0.28	Fever (38° or higher)	-0.27
Muscle pain/aches	0.27	Omicron variant	-0.26
Lower back pain	0.23	Heaviness in arms/legs	-0.08
Original variant	0.17	Very good health	-0.07
Feeling warm & cold	0.16	No chronic disease	-0.07
Red, painful eyes	0.16	Age group	-0.06
Sneezing	0.16	Smoker	-0.05
Difficulty breathing	0.14	Male	-0.03

Estimated coefficients of Linear Regression for prediction of post COVID-19 condition

Interpretation

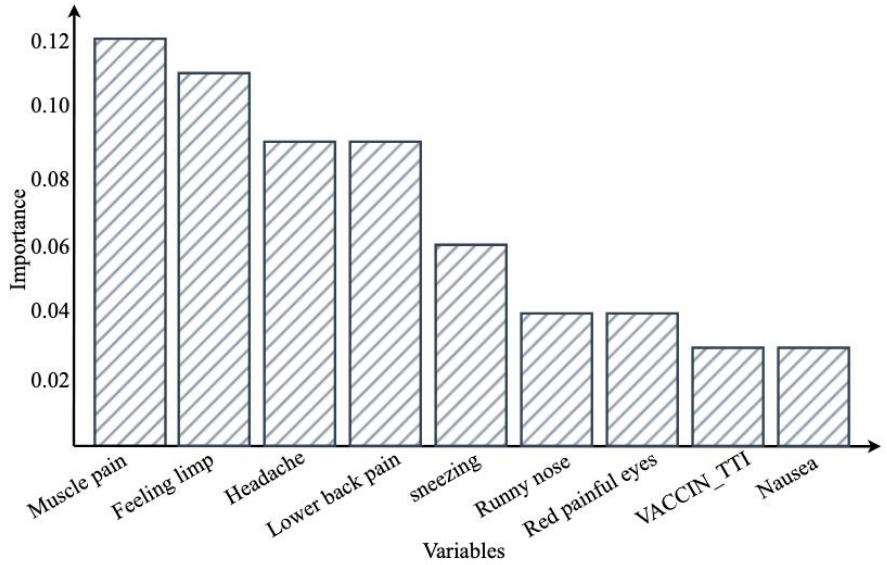
Random Forest

Top 10 features identified using **Random Forest**

- Shows some overlap with Linear Ridge Regression results, but **relative importance differs**

Key Predictors:

- **Muscle pain** → most important predictor of PCSI
- **Time between vaccination and infection (VACCIN_TTI)** → significant impact
- **Insight:**
  - Vaccination timing may influence **PCC risk and severity**
  - RF highlights **non-linear interactions** not captured by linear models



Feature importances resulted using Random Forest model for prediction of PCSI

Interpretation

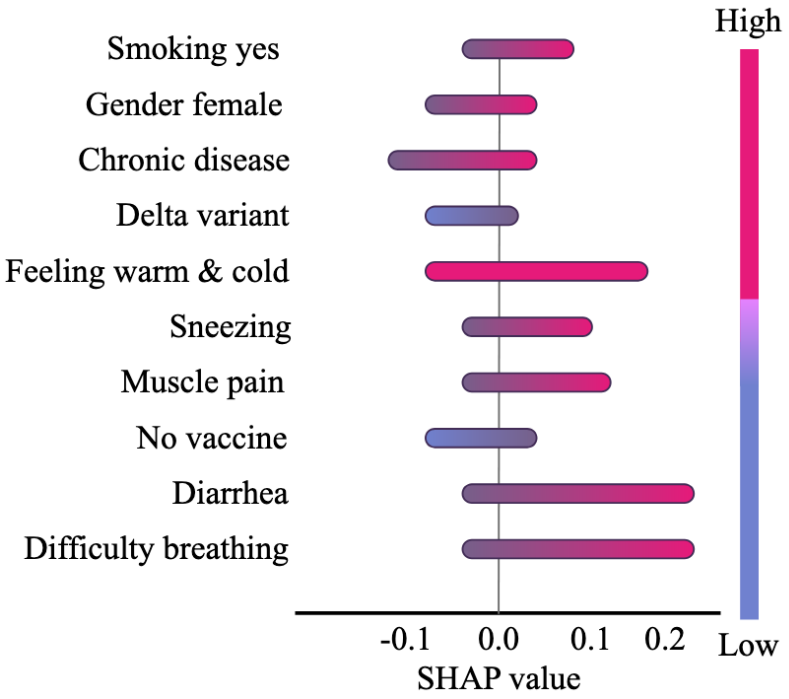
SHAP values

Top predictors for PCSI (via SHAP values):

- **Positive impact (↑ PCSI):** difficulty breathing, diarrhea, fluctuating body temperature, muscle pain, sneezing, smoking
- **Negative impact (↓ PCSI):** prior vaccination, absence of chronic diseases

Other insights:

- **Female sex** associated with higher PCSI
- Highlights the **complex interaction** of symptoms, baseline health, and demographics in PCC risk



Interpreting MLP influential factors using SHAP values

Conclusions and perspectives

Study Goal:

- Identify high-risk PCC profiles and predict **Post-COVID Symptom Intensity (PCSI)** using ML

Key Findings:

- **Higher risk:** women, patients with chronic diseases
- **Important predictors:** loss of smell, headache, muscle pain, vaccination timing
- **Protective factors:** absence of chronic diseases
- **MLP** slightly outperformed other models (lower MAPE)

Limitations:

- Steady-state assumption limits **temporal analysis**
- Dataset quality and completeness affect performance
- Models complement, **not replace**, clinical judgment

Societal Impact:

- PCC affects **health, daily life, and productivity**
- Predicting symptom intensity can guide **early interventions** and improve **patient outcomes**

17

References

[1] T. Lancet, The covid-19 pandemic in 2023: Far from over, 2023.

[2] A. V. Ballering, S. K. van Zon, T. C. olde Hartman, and J. G. Rosmalen, “Persistence of somatic symptoms after covid-19 in the netherlands: An observational cohort study,” The Lancet, vol. 400, no. 10350, pp. 452–461, 2022

[3] F. Callard and E. Perego, “How and why patients made long covid,” Social science & medicine, vol. 268, p. 113 426, 2021.

[4] https://www.who.int/europe/news-room/fact-sheets/item/post-COVID-19-condition. retrieved: September, 2025.

[5] C. E. Hastie et al., “Natural history of long-covid in a nationwide, population cohort study,” Nature Communications, vol. 14, no. 1, p. 3504, 2023.

[6] R. Kessler, J. Philipp, J. Wilfer, and K. Kostev, “Predictive attributes for developing long covid—a study using machine learning and real-world data from primary care physicians in germany,” Journal of Clinical Medicine, 2023.

Thank you for your attention!

Questions?