Identifying Post-COVID Risk Factors with Model-Agnostic Feature Importance

using the `xplainfi` R package with the German national cohort (NAKO)

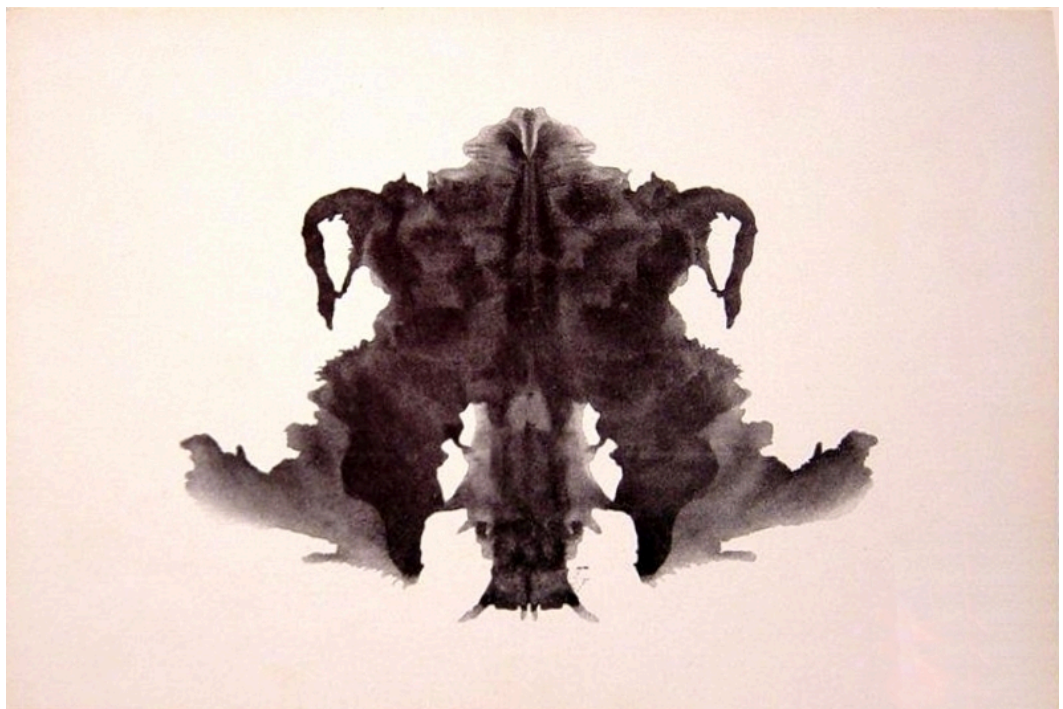
Lukas Burk

Leibniz Institute for Prevention Research and Epidemiology — BIPS
Bremen, Germany

2026-05-19

CEN 2026

A matter of interpretation



Post-COVID condition (PCC)

Not to be confused with “long COVID”

- Persistent / recurring symptoms long after SARS-CoV-2 infection
- Clinically heterogenous (symptoms & severity)

Post-COVID condition (PCC)

Not to be confused with “long COVID”

- Persistent / recurring symptoms long after SARS-CoV-2 infection
- Clinically heterogenous (symptoms & severity)
- Working definition: (Mikolajczyk et al. 2024)

Post-COVID condition (PCC)

Not to be confused with “long COVID”

- Persistent / recurring symptoms long after SARS-CoV-2 infection
- Clinically heterogenous (symptoms & severity)
- Working definition: (Mikolajczyk et al. 2024)
  - “Any” PCC: at least **1** symptom persistent **4+ months after** infection

Post-COVID condition (PCC)

Not to be confused with “long COVID”

- Persistent / recurring symptoms long after SARS-CoV-2 infection
- Clinically heterogenous (symptoms & severity)
- Working definition: (Mikolajczyk et al. 2024)
  - “Any” PCC: at least **1** symptom persistent **4+ months after** infection
  - “Severe” PCC: at least **9**

Post-COVID condition (PCC)

Not to be confused with “long COVID”

- Persistent / recurring symptoms long after SARS-CoV-2 infection
- Clinically heterogenous (symptoms & severity)
- Working definition: (Mikolajczyk et al. 2024)
  - “Any” PCC: at least **1** symptom persistent **4+ months after** infection
  - “Severe” PCC: at least **9**
- Many open questions: risk factors, causal factors, subgroups?

Post-COVID condition (PCC)

Not to be confused with “long COVID”

- Persistent / recurring symptoms long after SARS-CoV-2 infection
- Clinically heterogenous (symptoms & severity)
- Working definition: (Mikolajczyk et al. 2024)
  - “Any” PCC: at least **1** symptom persistent **4+ months after** infection
  - “Severe” PCC: at least **9**
- Many open questions: risk factors, causal factors, subgroups?

→ **RESOLVE-PCC** project funded by the German BMFTR

Post-COVID condition (PCC)

Not to be confused with “long COVID”

- Persistent / recurring symptoms long after SARS-CoV-2 infection
- Clinically heterogenous (symptoms & severity)
- Working definition: (Mikolajczyk et al. 2024)
  - “Any” PCC: at least **1** symptom persistent **4+ months after** infection
  - “Severe” PCC: at least **9**
- Many open questions: risk factors, causal factors, subgroups?

→ **RESOLVE-PCC** project funded by the German BMFTR

Federal Ministry of
Research, Technology and Space

Why feature importance?

What it can and can't tell us about PCC

- Predictive machine learning models use many candidate risk factors
- Which ones actually drive predictions?

Why feature importance?

What it can and can't tell us about PCC

- Predictive machine learning models use many candidate risk factors
- Which ones actually drive predictions?
- Applications:
  - Feature selection 🤝 feature importance

Why feature importance?

What it can and can't tell us about PCC

- Predictive machine learning models use many candidate risk factors
- Which ones actually drive predictions?
- Applications:
  - Feature selection 🤝 feature importance
  - Sanity check: is the model relying on plausible signal?

Why feature importance?

What it can and can't tell us about PCC

- Predictive machine learning models use many candidate risk factors
- Which ones actually drive predictions?
- Applications:
  - Feature selection 🤝 feature importance
  - Sanity check: is the model relying on plausible signal?
  - Hypothesis generation

Why feature importance?

What it can and can't tell us about PCC

- Predictive machine learning models use many candidate risk factors
- Which ones actually drive predictions?
- Applications:
  - Feature selection 🤝 feature importance
  - Sanity check: is the model relying on plausible signal?
  - Hypothesis generation
  - **FI $\neq$ causal effect** but associations still informative (Ewald et al. 2024)

Why feature importance?

What it can and can't tell us about PCC

- Predictive machine learning models use many candidate risk factors
- Which ones actually drive predictions?
- Applications:
  - Feature selection 🤝 feature importance
  - Sanity check: is the model relying on plausible signal?
  - Hypothesis generation
  - **FI $\neq$ causal effect** but associations still informative (Ewald et al. 2024)
- "Importance" itself is not a single quantity

Why feature importance?

What it can and can't tell us about PCC

- Predictive machine learning models use many candidate risk factors
- Which ones actually drive predictions?
- Applications:
  - Feature selection 🤝 feature importance
  - Sanity check: is the model relying on plausible signal?
  - Hypothesis generation
  - **FI $\neq$ causal effect** but associations still informative (Ewald et al. 2024)
- "Importance" itself is not a single quantity → *method choice* defines the answer

Common FI methods

How “important” is feature X_j for prediction?

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance
4. (Repeat k times for stability)

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance
4. (Repeat k times for stability)

Leave-one-covariate-out (*LOCO*)

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance
4. (Repeat k times for stability)

Perturbation of X_j

1. Fit full model
2. Measure performance on...

Leave-one-covariate-out (*LOCO*)

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance
4. (Repeat k times for stability)

Perturbation of X_j

1. Fit full model
2. Measure performance on...
   - a) all test data

Leave-one-covariate-out (*LOCO*)

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance
4. (Repeat k times for stability)

Perturbation of X_j

1. Fit full model
2. Measure performance on...
   - a) all test data
   - b) same data where X_j is *randomly permuted*
3. Compare performance

Leave-one-covariate-out (*LOCO*)

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance
4. (Repeat k times for stability)

Perturbation of X_j

1. Fit full model
2. Measure performance on...
   - a) all test data
   - b) same data where X_j is *randomly permuted*
3. Compare performance
4. (Repeat k times for stability)

Leave-one-covariate-out (*LOCO*)

Common FI methods

How "important" is feature X_j for prediction?

Refitting without feature X_j

1. Fit full model
2. Fit model without X_j
3. Compare performance
4. (Repeat k times for stability)

Perturbation of X_j

1. Fit full model
2. Measure performance on...
   - a) all test data
   - b) same data where X_j is *randomly permuted*
3. Compare performance
4. (Repeat k times for stability)

Leave-one-covariate-out (*LOCO*)

Permutation feature importance (*PFI*)

Conditional FI

Permute, but respect feature dependence

- PFI permutes X_j marginally → implausible combinations (Hooker et al. 2021)
- E.g.: “20-year-old with 30 years smoking history”

Conditional FI

Permute, but respect feature dependence

- PFI permutes X_j marginally → implausible combinations (Hooker et al. 2021)
- E.g.: “20-year-old with 30 years smoking history”
- CFI: perturbation **conditional** on other features

Conditional FI

Permute, but respect feature dependence

- PFI permutes X_j marginally → implausible combinations (Hooker et al. 2021)
- E.g.: “20-year-old with 30 years smoking history”
- CFI: perturbation **conditional** on other features
- Requires *conditional sampling* $\tilde{X}_j \sim F_{X_j \mid X_{-j}}$, some options:

Conditional FI

Permute, but respect feature dependence

- PFI permutes X_j marginally → implausible combinations (Hooker et al. 2021)
- E.g.: “20-year-old with 30 years smoking history”
- CFI: perturbation **conditional** on other features
- Requires *conditional sampling* $\tilde{X}_j \sim F_{X_j \mid X_{-j}}$, some options:
  - Conditional Gaussian → fast, but only continuous data

Conditional FI

Permute, but respect feature dependence

- PFI permutes X_j marginally → implausible combinations (Hooker et al. 2021)
- E.g.: “20-year-old with 30 years smoking history”
- CFI: perturbation **conditional** on other features
- Requires *conditional sampling* $\tilde{X}_j \sim F_{X_j \mid X_{-j}}$, some options:
  - Conditional Gaussian → fast, but only continuous data
  - Adversarial random forest (ARF) (Blesch et al. 2025)

Conditional FI

Permute, but respect feature dependence

- PFI permutes X_j marginally → implausible combinations (Hooker et al. 2021)
- E.g.: “20-year-old with 30 years smoking history”
- CFI: perturbation **conditional** on other features
- Requires *conditional sampling* $\tilde{X}_j \sim F_{X_j \mid X_{-j}}$, some options:
  - Conditional Gaussian → fast, but only continuous data
  - Adversarial random forest (ARF) (Blesch et al. 2025)

    → handles *mixed* data, *missing values*, computationally more expensive

What are we explaining?

Fixed model, learning algorithm, data-generating process?

- **Model-level**: explain a single fitted model
  - One or many permutations (**PFI/CFI**) on a held-out set

What are we explaining?

Fixed model, learning algorithm, data-generating process?

- **Model-level**: explain a single fitted model
  - One or many permutations (**PFI/CFI**) on a held-out set
- **Learner-level FI**: explain the *prediction method*, not one fit
  - **LOCO** automatically refits, includes learner variability

What are we explaining?

Fixed model, learning algorithm, data-generating process?

- **Model-level**: explain a single fitted model
  - One or many permutations (**PFI/CFI**) on a held-out set
- **Learner-level FI**: explain the *prediction method*, not one fit
  - **LOCO** automatically refits, includes learner variability
  - **Any method**: Repeat across resampling (e.g. bootstrap, subsampling, CV)
  - Captures variability from data sampling and learner stochasticity

Analysis dataset: NAKO

German National Cohort

- $N \approx 66.000$ participants reporting ≥ 1 SARS-CoV-2 infection
- $\approx 24\%$ classified as “any PCC” (“Severe PCC” much rarer → not shown today)

Analysis dataset: NAKO

German National Cohort

- $N \approx 66.000$ participants reporting ≥ 1 SARS-CoV-2 infection
- $\approx 24\%$ classified as “any PCC” (“Severe PCC” much rarer → not shown today)
- → **Goal**: Predict “any PCC” vs. “no PCC” (binary classif)

Analysis dataset: NAKO

German National Cohort

- $N \approx 66.000$ participants reporting ≥ 1 SARS-CoV-2 infection
- $\approx 24\%$ classified as “any PCC” (“Severe PCC” much rarer → not shown today)
- → **Goal**: Predict “any PCC” vs. “no PCC” (binary classif)
- Mix of *survey* + *in-person assessment*:
  - demographics, anthropometrics, socioeconomic
  - comorbidities, smoking history, biomarker labs (where available)
  - mental-health questionnaires

Analysis dataset: NAKO

German National Cohort

- $N \approx 66.000$ participants reporting ≥ 1 SARS-CoV-2 infection
- $\approx 24\%$ classified as “any PCC” (“Severe PCC” much rarer → not shown today)
- → **Goal**: Predict “any PCC” vs. “no PCC” (binary classif)
- Mix of *survey* + *in-person assessment*:
  - demographics, anthropometrics, socioeconomic
  - comorbidities, smoking history, biomarker labs (where available)
  - mental-health questionnaires
- $\approx$ 50 features used

Analysis: ‘Any PCC’ prediction

Preliminary, exploratory

- Learners: Gradient boosting (`XGBoost`), random forest (`ranger`)
- Tuned on **PR-AUC** (baseline = 24%)

Analysis: ‘Any PCC’ prediction

Preliminary, exploratory

- Learners: Gradient boosting (`XGBoost`), random forest (`ranger`)
- Tuned on **PR-AUC** (baseline = 24%)
- General performance:
  - PR-AUC $\approx$ 43%
  - ROC-AUC $\approx$ 64%

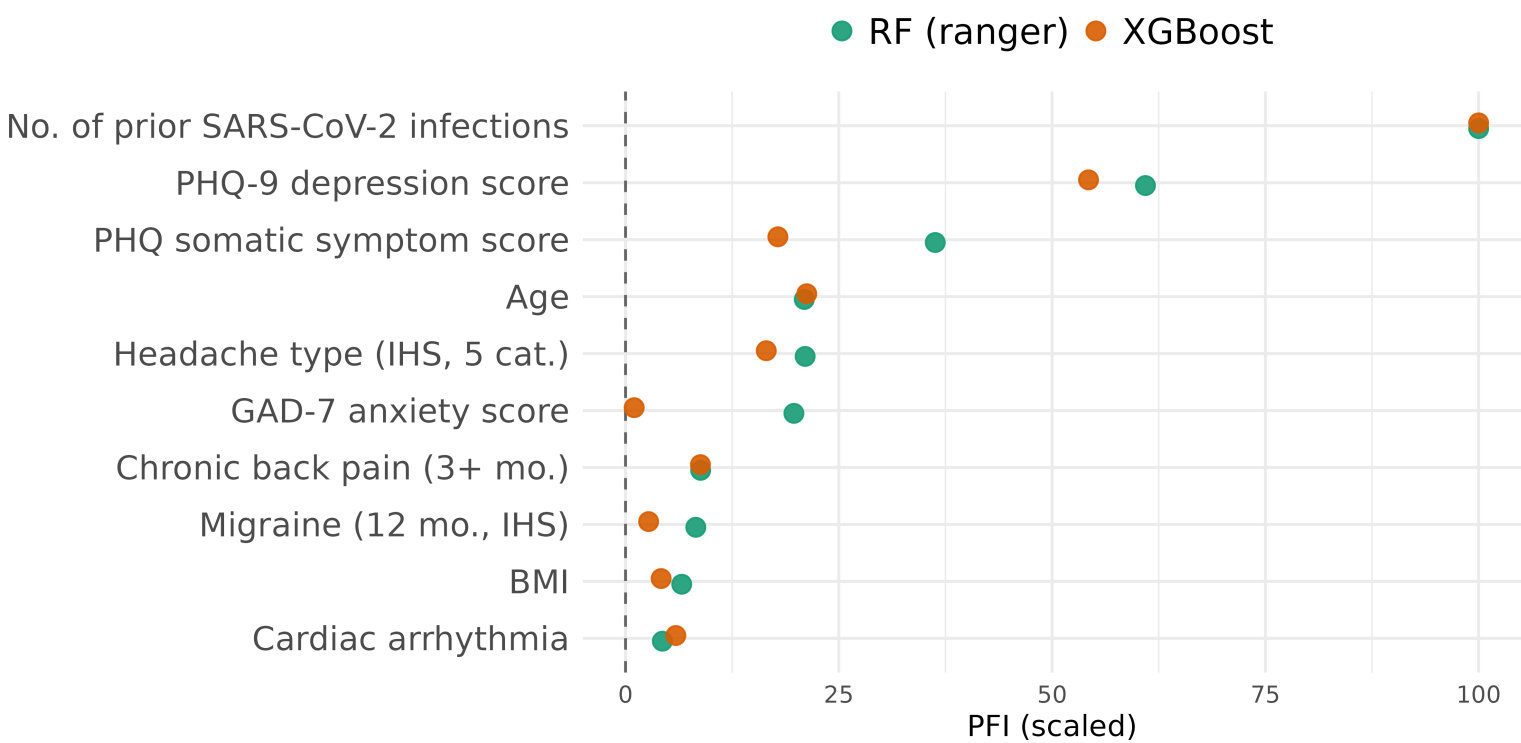
Analysis: ‘Any PCC’ prediction

Preliminary, exploratory

- Learners: Gradient boosting (`XGBoost`), random forest (`ranger`)
- Tuned on **PR-AUC** (baseline = 24%)
- General performance:
  - PR-AUC $\approx$ 43%
  - ROC-AUC $\approx$ 64%
- **PFI**, **CFI** (+ARF), **LOCO** computed on *test set* (model importance)

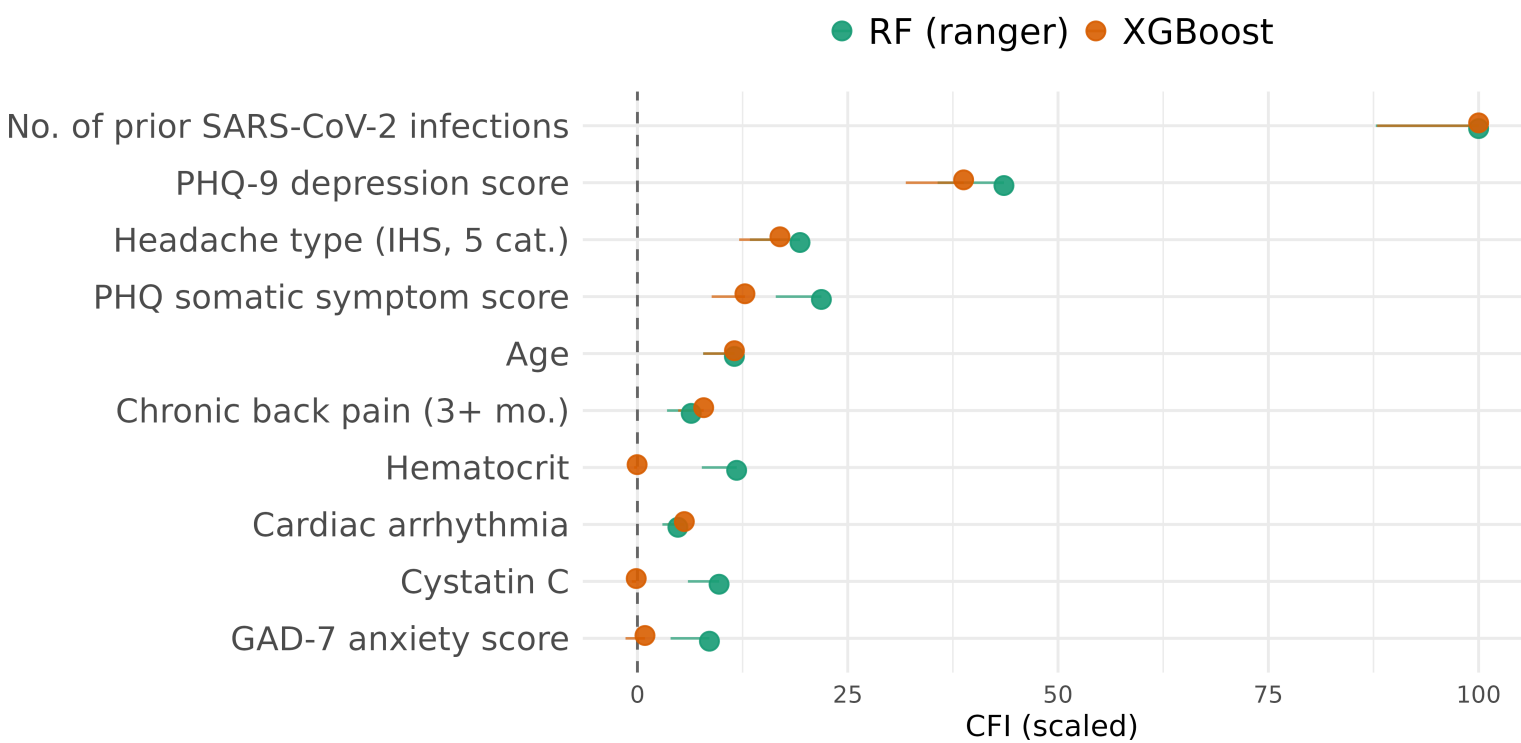
Results: Permutation Feature Importance (PFI)

Top 10 features (between RF + XGBoost)



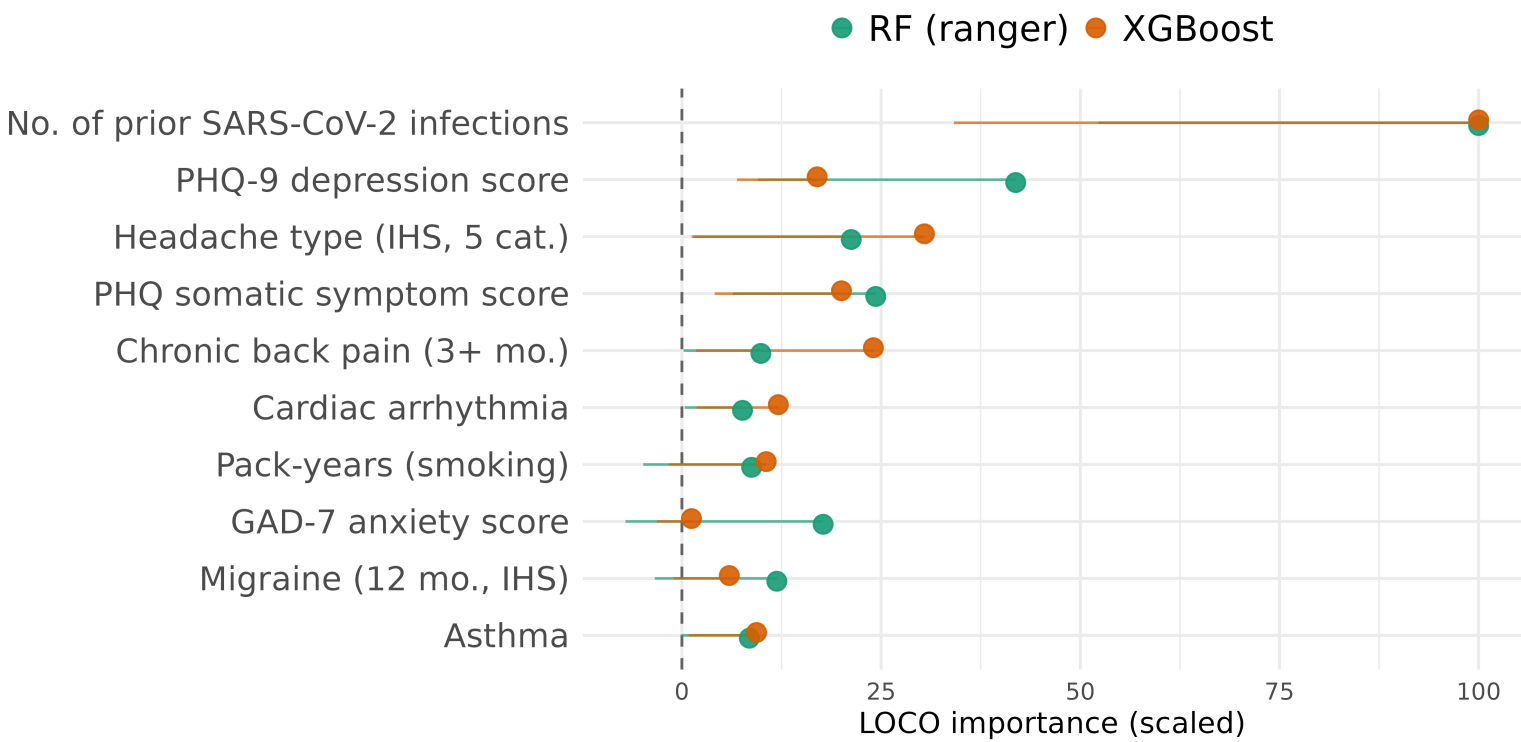
Results: Conditional Feature Importance (CFI)

Top 10 features, one-sided 95% CIs



Results: Leave-one-covariate-out (LOCO)

Top 10 features, one-sided 95% CIs



Conclusion

Everything is complicated, always

- Feature importance is useful, but **not a single number**
- Each method answers a slightly different question

Conclusion

Everything is complicated, always

- Feature importance is useful, but **not a single number**
- Each method answers a slightly different question
  - **PFI**: marginal, model-faithful but extrapolation issue

Conclusion

Everything is complicated, always

- Feature importance is useful, but **not a single number**
- Each method answers a slightly different question
  - **PFI**: marginal, model-faithful but extrapolation issue
  - **CFI**: conditional, more faithful to joint distr., sampling-dependent

Conclusion

Everything is complicated, always

- Feature importance is useful, but **not a single number**
- Each method answers a slightly different question
  - **PFI**: marginal, model-faithful but extrapolation issue
  - **CFI**: conditional, more faithful to joint distr., sampling-dependent
  - **LOCO**: refit-based, expensive but assumption-light

Conclusion

Everything is complicated, always

- Feature importance is useful, but **not a single number**
- Each method answers a slightly different question
  - **PFI**: marginal, model-faithful but extrapolation issue
  - **CFI**: conditional, more faithful to joint distr., sampling-dependent
  - **LOCO**: refit-based, expensive but assumption-light
- Don’t forget the role of the **learner**
- Compare methods → robustness check, not contradiction

Conclusion

Everything is complicated, always

- Feature importance is useful, but **not a single number**
- Each method answers a slightly different question
  - **PFI**: marginal, model-faithful but extrapolation issue
  - **CFI**: conditional, more faithful to joint distr., sampling-dependent
  - **LOCO**: refit-based, expensive but assumption-light
- Don’t forget the role of the **learner**
- Compare methods → robustness check, not contradiction
- `xplainfi` provides a unified interface for all of the above

Conclusion

Everything is complicated, always

- Feature importance is useful, but **not a single number**
- Each method answers a slightly different question
  - **PFI**: marginal, model-faithful but extrapolation issue
  - **CFI**: conditional, more faithful to joint distr., sampling-dependent
  - **LOCO**: refit-based, expensive but assumption-light
- Don’t forget the role of the **learner**
- Compare methods → robustness check, not contradiction
- `xplainfi` provides a unified interface for all of the above (Burk et al. 2026)

Thank you for your attention!

Contact

Lukas Burk
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
28359 Bremen
Germany

burk@leibniz-bips.de

References

Appel KS, Nürnberger C, Bahmer T, et al (2024) Definition of the Post-COVID Syndrome Using a Symptom-Based Post-COVID Score in a Prospective, Multi-Center, Cross-Sectoral Cohort of the German National Pandemic Cohort Network (NAPKON). Infection 52:1813–1829. https://doi.org/10.1007/s15010-024-02226-9

Blesch K, Koenen N, Kapar J, et al (2025) Conditional Feature Importance with Generative Modeling Using Adversarial Random Forests. Proceedings of the AAAI Conference on Artificial Intelligence 39:15596–15604. https://doi.org/10.1609/aaai.v39i15.33712

Burk L, Ewald FK, Casalicchio G, et al (2026) Xplainfi: Feature Importance and Statistical Inference for Machine Learning in R

Ewald FK, Bothmann L, Wright MN, et al (2024) A Guide to~Feature Importance Methods for~Scientific Inference. In: Longo L, Lapuschkin S, Seifert C (eds) Explainable Artificial Intelligence. Springer Nature Switzerland, Cham, pp 440–464

Hooker G, Mentch L, Zhou S (2021) Unrestricted Permutation Forces Extrapolation: Variable Importance Requires at Least One More Model, or There Is No Free Variable Importance. Statistics and Computing 31:82. https://doi.org/10.1007/s11222-021-10057-z

Mikolajczyk R, Diexer S, Klee B, et al (2024) Likelihood of Post-COVID Condition in People with Hybrid Immunity; Data from the German National Cohort (NAKO). The Journal of Infection 89:106206. https://doi.org/10.1016/j.jinf.2024.106206