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

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

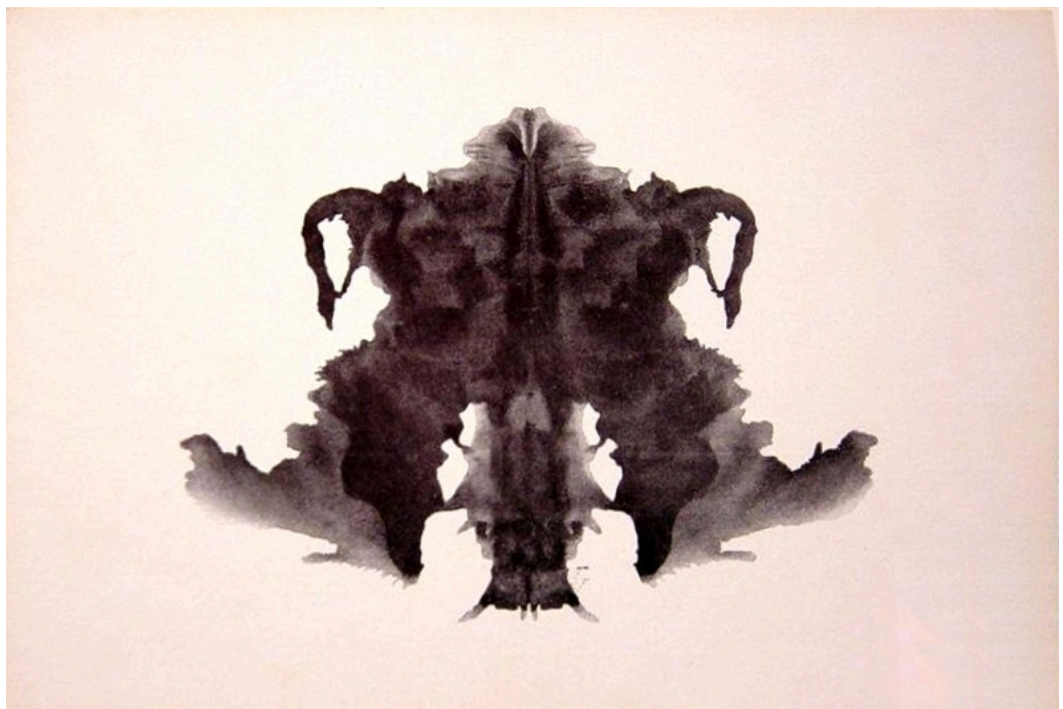
Lukas Burk

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

2026-05-19

CEN 2026

A matter of interpretation



Post-COVID condition (PCC)

“Long- / Post-COVID”

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

Post-COVID condition (PCC)

“Long- / Post-COVID”

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

Post-COVID condition (PCC)

“Long- / Post-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**, “severe” PCC: at least **9**
  - … symptoms persistent **4+ months after** infection

Post-COVID condition (PCC)

“Long- / Post-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**, “severe” PCC: at least **9**
  - … symptoms persistent **4+ months after** infection
- Many open questions: risk factors, causal factors, subgroups?

Post-COVID condition (PCC)

“Long- / Post-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**, “severe” PCC: at least **9**
  - … symptoms persistent **4+ months after** infection
- Many open questions: risk factors, causal factors, subgroups?

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

Post-COVID condition (PCC)

“Long- / Post-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**, “severe” PCC: at least **9**
  - … symptoms persistent **4+ months after** infection
- 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 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 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 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 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 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 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?

Common FI methods

How "important" is feature X_j?

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?

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?

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?

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?

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?

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?

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?

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 forests (ARF)

    → 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. 15 subsampling iters)
  - Captures variability from data sampling and learner stochasticity

Analysis dataset: NAKO

German National Cohort

- $N \approx 66250$ 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

- $N \approx 66250$ participants reporting ≥ 1 SARS-CoV-2 infection
- $\approx 24\%$ classified as **“any PCC”** (“Severe PCC” much rarer → not shown today)
- 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

- $N \approx 66250$ participants reporting ≥ 1 SARS-CoV-2 infection
- $\approx 24\%$ classified as **“any PCC”** (“Severe PCC” much rarer → not shown today)
- Mix of *survey* + *in-person assessment*:
  - demographics, anthropometrics, socioeconomic
  - comorbidities, smoking history, biomarker labs (where available)
  - mental-health questionnaires
- ≈ 150 features used (dropping near-constant / near-fully-missing)
- Rare-event indicators retained despite low prevalence

Analysis: ‘Any PCC’ prediction

Preliminary, exploratory

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

Analysis: ‘Any PCC’ prediction

Preliminary, exploratory

- Learners: Gradient boosting (`XGBoost`), random forest (`ranger`)
- Tuned on **PR-AUC** (due to imbalance, baseline = 24%)
  - → general performance **42–44% PR-AUC**
- **PFI**, **CFI** (+ARF), **LOCO** computed via `xplainfi`

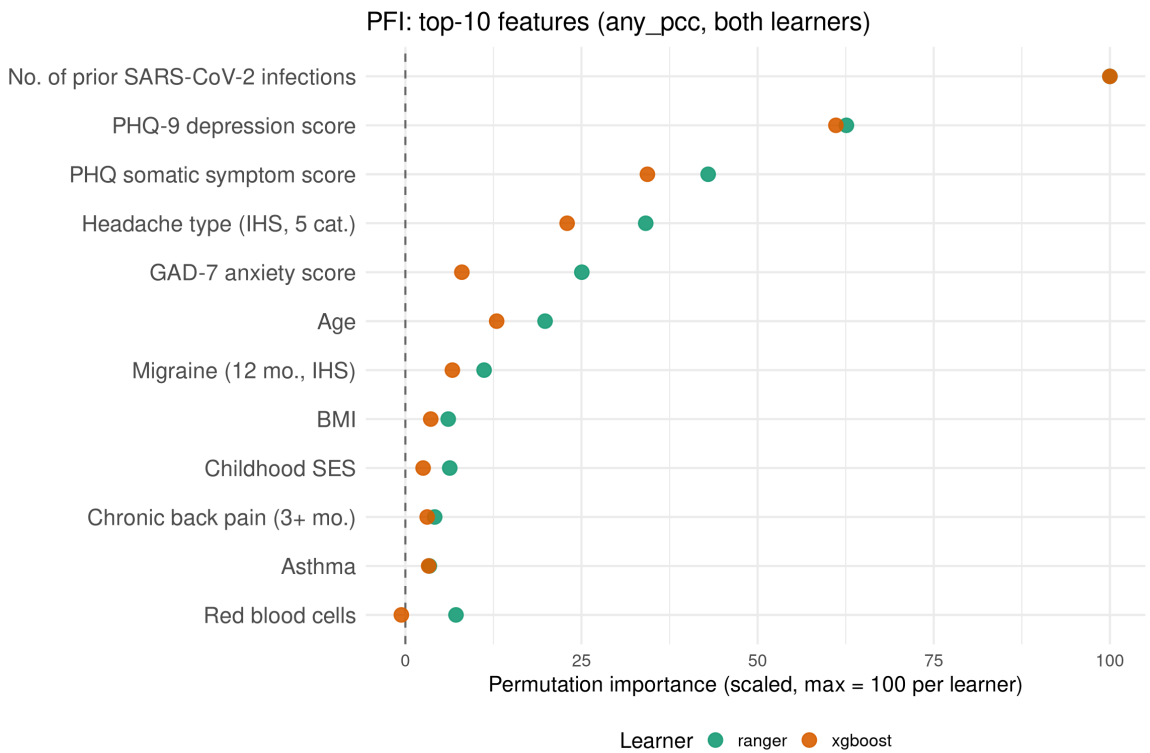
Analysis: ‘Any PCC’ prediction

Preliminary, exploratory

- Learners: Gradient boosting (`XGBoost`), random forest (`ranger`)
- Tuned on **PR-AUC** (due to imbalance, baseline = 24%)
  - → general performance **42–44% PR-AUC**
- **PFI**, **CFI** (+ARF), **LOCO** computed via `xplainfi`
- FI here on test set (i.e. **model importance**)

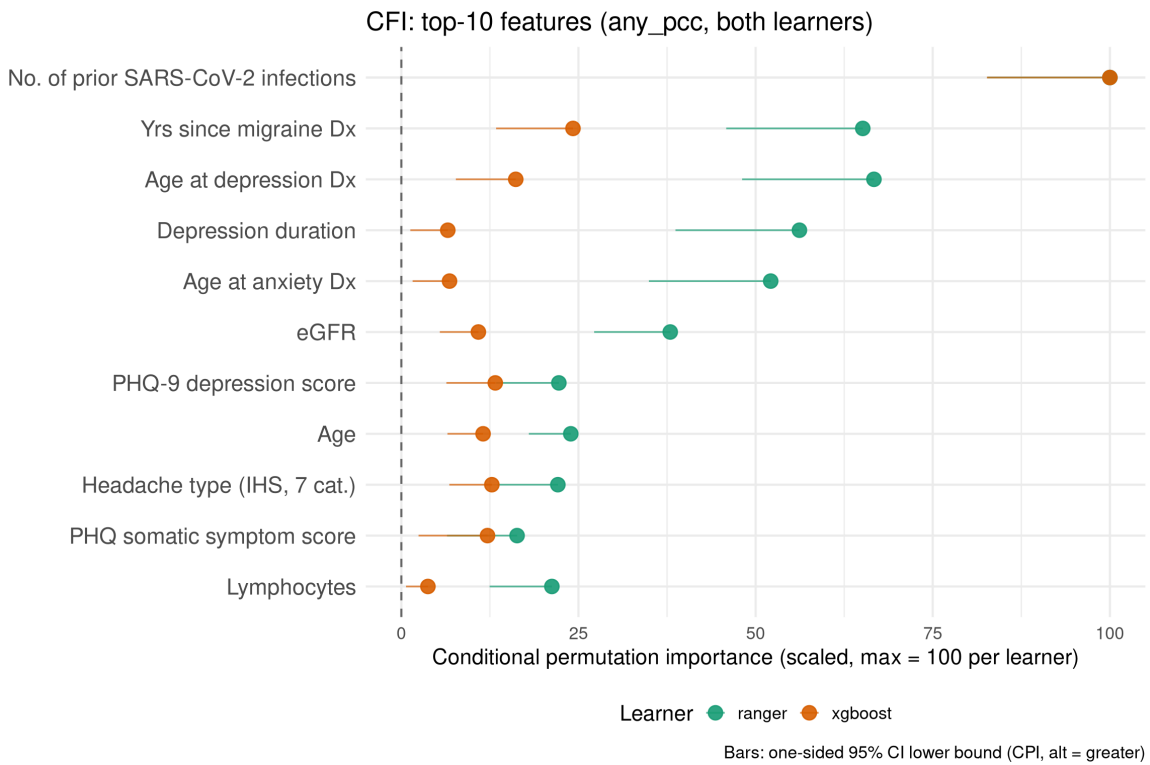
Results: PFI

Top 10 features



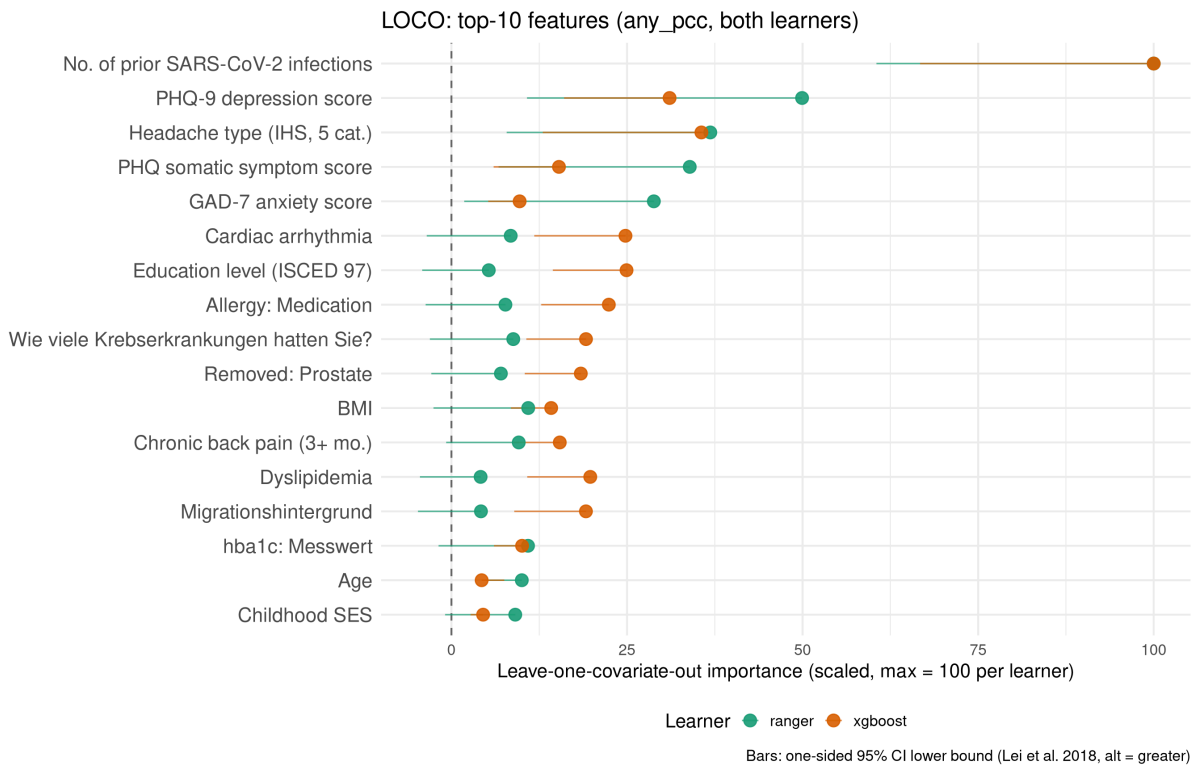
Results: CFI

Top 10 features



Results: LOCO

Top 10 features



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
- Match the **estimand** to the **research question**
- 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, K. S., Nürnberger, C., Bahmer, T., Förster, C., Polidori, M. C., Kohls, M., Kraus, T., Hettich-Damm, N., Petersen, J., Blaschke, S., Bröhl, I., Butzmann, J., Dashti, H., Deckert, J., Dreher, M., Fiedler, K., Finke, C., Geisler, R., Hanses, F., … NAPKON Study Group. (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*(5), 1813–1829. https://doi.org/10.1007/s15010-024-02226-9

Burk, L., Ewald, F. K., Casalicchio, G., Wright, M. N., & Bischl, B. (2026, March 16). *Xplainfi: Feature Importance and Statistical Inference for Machine Learning in R*. https://doi.org/10.48550/arXiv.2603.15306

Ewald, F. K., Bothmann, L., Wright, M. N., Bischl, B., Casalicchio, G., & König, G. (2024). A Guide to~Feature Importance Methods for~Scientific Inference. In L. Longo, S. Lapuschkin, & C. Seifert (Eds.), *Explainable Artificial Intelligence: Explainable Artificial Intelligence*. https://doi.org/10.1007/978-3-031-63797-1_22

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*(6), 82. https://doi.org/10.1007/s11222-021-10057-z

Mikolajczyk, R., Diexer, S., Klee, B., Pfrommer, L., Purschke, O., Fricke, J., Ahnert, P., Gabrysch, S., Gottschick, C., Bohn, B., Brenner, H., Buck, C., Castell, S., Gastell, S., Greiser, K. H., Harth, V., Heise, J.-K., Holleczek, B., Kaaks, R., … Karch, A. (2024). Likelihood of Post-COVID Condition in People with Hybrid Immunity; Data from the German National Cohort (NAKO). *The Journal of Infection*, *89*(2), 106206. https://doi.org/10.1016/j.jinf.2024.106206