Why predicting risk can't identify 'risk factors': empirical assessment of model stability in machine learning across observational health databases

PRESENTER: Aniek Markus (a.markus@erasmusmc.nl)
CO-AUTHORS: Peter R. Rijnbeek, Jenna M. Reps

INTRODUCTION:

- Some researchers incorrectly interpret prediction models:
  - Interpreting selected variables as factors that cause the outcome
  - **Using selected variables for `risk factor` detection (i.e. to identify variables associated with the outcome)**
- We illustrate potential issues by investigating the stability of >450 prediction models in a large-scale experiment, investigating model changes across databases (care settings) and phenotype definitions.

METHODS:

1. We developed models using LASSO logistic regression for nine prediction tasks: predicting nine COVID-19 vaccine outcomes of interest (O) identified by the U.S. Food and Drug Administration for the general population (T) in the next 1 year (TAR).
2. Measure model stability:

Q1. How many variables are selected across models?
Q2. Are the same or different variables included across models?
Q3. Is the direction of the effect of variables the same across models?

RESULTS:

- Q1: A higher number of outcome cases generally leads to more variables being selected using (Fig 3).
- Q2: Overall model stability was poor, slightly better for top (i.e. most important) variables (Fig 4). The impact of different target/outcome phenotype definitions was limited, but the top 10 variables differed across databases (Fig 1).
- Q3: The sign of the coefficient can vary greatly even for the top variables (Fig 2), less selected variables seem more likely to switch sign.

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Be careful interpreting prediction models as the identified 'risk factors' appear to depend on study design choices.

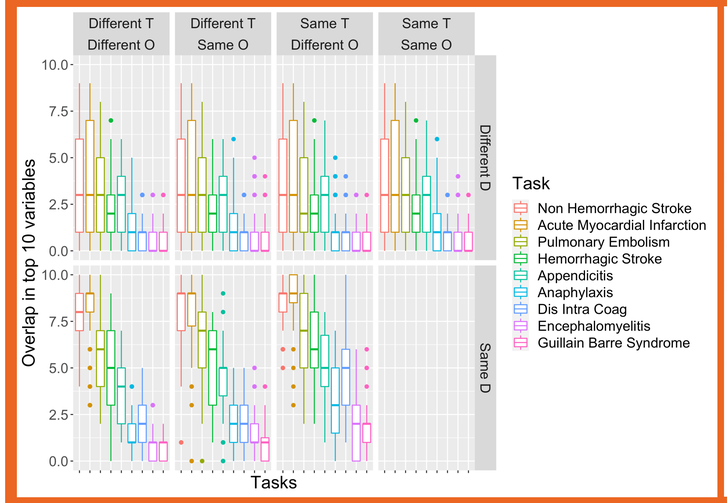


Figure 1. Overlap in the top 10 variables as defined by counting the number of common variables between each pair of models for same/different database (D), target population definition (T), outcome definition (O) across models

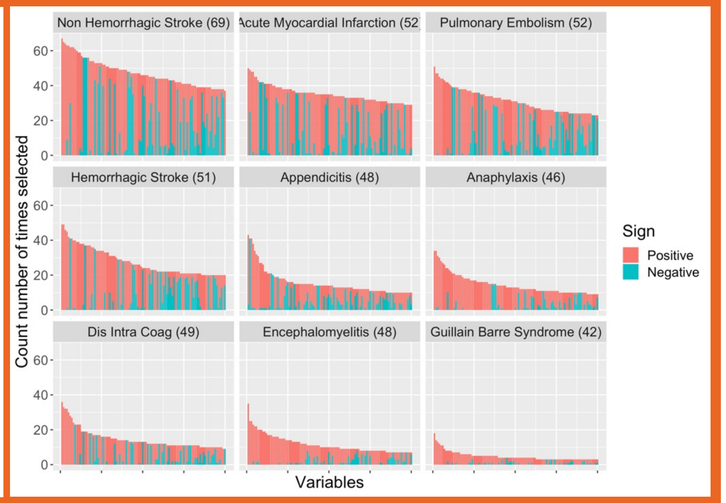


Figure 2. Graph visualizing the number of times variables are selected and the percentage of times these variables had a positive or negative sign for each prediction task.

TAKE AWAYS:

- There is substantial variation in the selected variables across models.
- Different databases lead to different 'risk factors'.
- Interpreting the effect of 'risk factors' is problematic as the sign can differ across models.
- We recommend investigating model robustness across settings or using other techniques for 'risk factor' detection (e.g. univariate analysis).

T	O	Databases
General population	Acute myocardial infarction, Anaphylaxis, Appendicitis, Disseminated intravascular coagulation, Encephalomyelitis, Guillain-Barré syndrome, Hemorrhagic stroke, Non-hemorrhagic stroke, Pulmonary embolism	CCAE, Optum EHR, Optum DoD, MDCD, IQVIA Germany, JMDC, MDCR

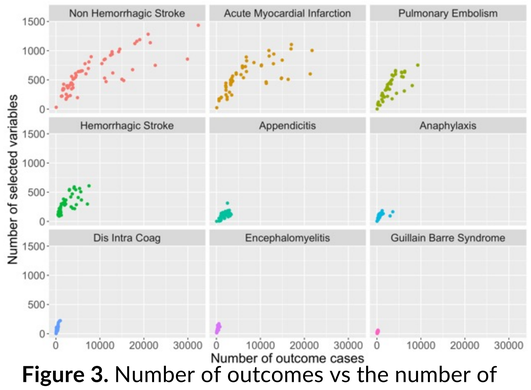
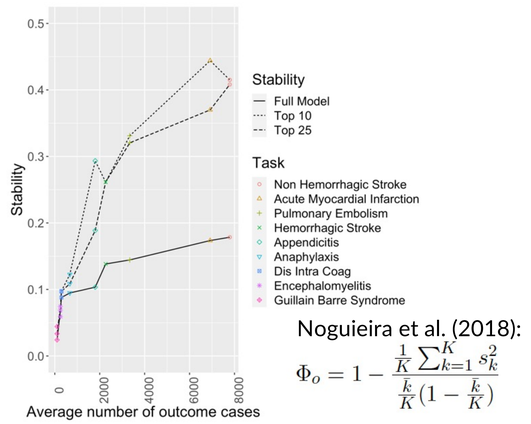


Figure 3. Number of outcomes vs the number of selected variables per prediction task.



Noguieira et al. (2018):

$$\Phi_o = 1 - \frac{\frac{1}{K}\sum_{k=1}^{K} s_k^2}{\frac{k}{K}(1-\frac{k}{K})}$$

Figure 4. Model stability (stability estimator Noguieira et al. 2018) vs the average number of outcome cases across prediction tasks.