Best practices for prognostic model development using observational health data: a scoping review

Cynthia Yang
PhD Student at Erasmus MC
Prognostic model development

- Among a **target population** of patients at an index date \( t = 0 \), predict which patients will experience some **outcome** during a **time-at-risk** period.
Study design choices

- Four study design components of prognostic model development:
Scoping review

• Aim: to present an overview of current knowledge and empirical evidence.

• Preliminary PubMed search for empirical studies.
  1. 6,671 titles inspected of studies from January 2018 to June 2022.
  2. 36 potentially relevant studies, 11 papers selected.
  3. Data extraction: study design component, recommendations for best practices, whether the OMOP CDM was used, and classifier.
Preliminary results

Data extraction
- Phenotypes
- Covariates
- Lookback period

Data pre-processing
- Observation time
- Loss to follow-up
- Sample size
- Missing data

Model development
- Classifier
- Hyperparameters
- Class imbalance
- Ensemble learning

Model validation
- Validation strategy
- Evaluation measures
- Recalibration
- Model updating

A short lookback (<180 days) can limit a model’s performance. A one-year lookback provides a good trade-off between performance and interpretability.

Excluding patients without the outcome who are lost to follow-up while including patients with the outcome who are then lost to follow-up leads to model bias.

More complex models (gradient boosting, random forest) had no consistent advantages over simpler ones (logistic regression).

To limit overfitting: (1) use a hold-out set to pick any hyperparameters, and (2) use a hold-out set to evaluate the model internally.
Preliminary results

- 6 out of 11 studies used OMOP CDM data.
- 3 out of 11 studies used MIMIC-III data.
- 2 out of 11 studies used other data.
- 5 out of 11 studies only used logistic regression.
- 6 out of 11 studies also used other classifiers.
Conclusions

- Model development should follow learned best practices.
- A lot of study design choices still require more research.
- Next steps: improve search strategy, more empirical research!
- For more information or questions please visit me at my poster.