How much data do we need for clinical prediction?

PRESENTER: **Henrik** John

INTRO

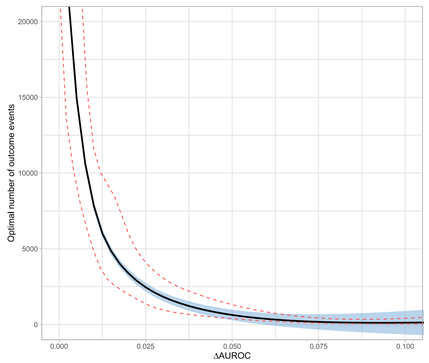
- **Large observational health databases are widely available, for example** IBM® MarketScan® Research Databases, which provide de-identified longitudinal patient-level data of more than 245 million individuals in the OMOP CDM.
- **OHDSI has developed a powerful prediction framework:** the patient-level prediction R-package.
- Our research question: **Does this amount of data help to improve prediction performance** or does it unnecessarily put constraints on computing resources and computation time?

METHODS

- We assessed the effect of sample size on prediction performance using **learning curves.**
- We generated learning curves for 81 prediction problems in three databases **building about 60,000 prediction models in the process.**

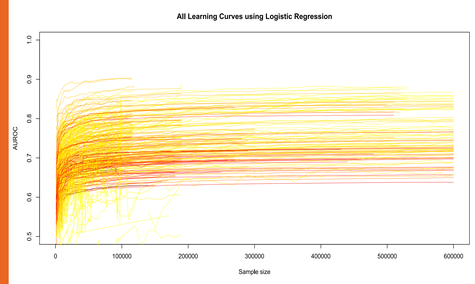
RESULTS

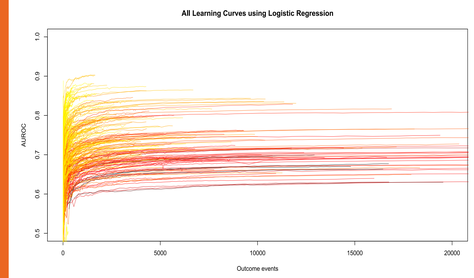
- We should ensure sufficient outcome events rather than a large enough sample size.
- We can empirically estimate an optimal number of outcome events given the graph below. The performance increase beyond 10,000 outcome events may be negligible for common application.

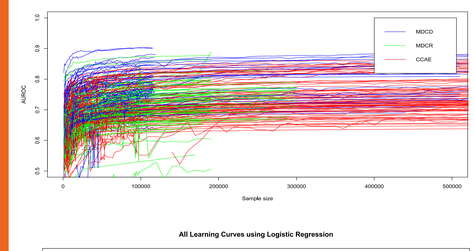


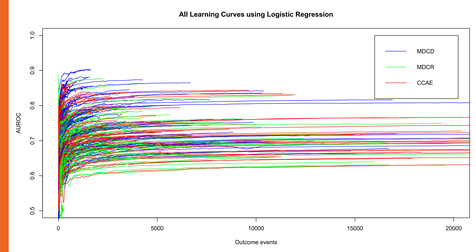
Optimal **sample size** for prediction should be defined in terms of the number of **outcome events**; and we **do not** need data from **millions** of patients.

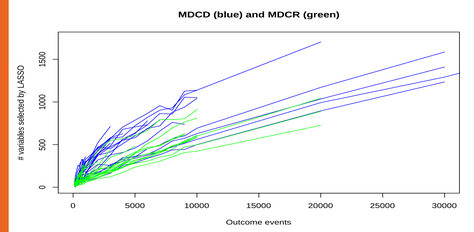
Take a picture to **download** the **full poster**











L.H. John, P.R. Rijnbeek
l.john@erasmusmc.nl

Erasmus MC Universitair Medisch Centrum Rotterdam

OHDSI