IBS-DR 2021

Education for Statistics in Practice

Development and evaluation of prediction models: pitfalls and solutions

Presenters: Ben Van Calster[1] & Maarten van Smeden[2]

[1]Department of Development and Regeneration, KU Leuven, Leuven, Belgium & department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands.
ben.vancalster@kuleuven.be
[2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, Netherlands.
m.vansmeden@umcutrecht.nl

EXTENDED ABSTRACT

Prediction models are developed throughout science. In this session the focus will be on applications in the medical domain, where prediction models have a long history commonly serving either a diagnostic or prognostic purpose. The ultimate goal of such models is to assist in medical decision making by providing accurate predictions for future individuals.

As we anticipate participants to this session are already well versed in fitting statistical models to data, the focus will be on the common pitfalls when developing statistical (learning) and machine learning models with a prediction aim. Our goal is that the participants gain knowledge about the pitfalls of prediction modeling and increase their familiarity with methods providing solutions for these pitfalls.

This session will be arranged in sections of 20 to 30 minutes. The following topics will be covered.

State of the medical prediction modeling art
This section begins with a small introduction into the history of prediction modeling in medical research. Positive examples will be highlighted and we will draw from the extensive systematic review literature on clinical prediction models. Recent experiences with a living systematic review on COVID-19 related prediction modeling will be discussed.

Just another prediction model
For most health conditions prediction models already exist. How does one prevent that prediction modeling project ends up on the large failed and unused model pile? Using the PROGRESS framework, we discuss various prediction modeling goals. Some good modeling practices and the harm of commonly applied modeling methods are illustrated. Finally, we will highlight some recent developments in formalizing prediction goals (*predictimands*).

Methods against overfitting
Overfitting is arguably the biggest enemy of prediction modeling. There is a large literature on shrinkage estimators that aim at preventing overfitting. In this section we will reflect on the history of shrinkage methods (e.g. Stein's estimator & Le Cessie van Houwelingen heuristic shrinkage) and more recent developments (e.g. lasso and ridge regression variants). The advantages and limitations will be discussed.

Methods for deciding on appropriate sample size

Rules of thumb have dominated the discussions on sample size for prediction models for decades (e.g. the need for at least 10 events for every predictor considered). The history and limitations of these rules of thumb will be shown. Recently developed sample size criteria for prediction model development and validation will be presented.

Model performance and validation
Validation of prediction models goes beyond evaluation of model coefficients and goodness-of-fit tests. Prediction models should give higher risk estimates for events than for non-events (discrimination). Predictions may be used to support clinical decisions, therefore the risks should be accurate (calibration). We will describe various levels at which a model can be calibrated. Further, the performance of the model to classify patients into low vs high risk patients to support decision making can be evaluated. We discuss decision curve analysis, the most well-known tool for utility validation. The link between calibration and utility is explained.

Heterogeneity over time and place: there is no such thing as a validated model
We discuss the different levels of validation (apparent, internal, and external), and what they can tell us. However, it is increasingly recognized that one should expect performance to be heterogeneous between different settings/hospitals. This can be taken into account on many levels: we may focus on having clustered (e.g. multicenter data, IPD) datasets for model development and validation, internal-external cross-validation can be used during model development, and cluster-specific performance can be meta-analyzed at validation. If data allow, meta-regression can be used to gain insight in performance heterogeneity. Model updating can be used to adapt a model to a new setting. In addition, populations tend to change over time. This calls for continuous updating strategies.

Applied example
We will describe the development and validation of the ADNEX model to diagnose ovarian cancer. Development, validation, target population, meta-regression, validation studies, model updating, and implementation in ultrasound machines.

Future perspective: machine learning and AI
Flexible machine learning algorithms have been around for a while. Recently, however, we have observed a strong increase in their use. We discuss challenges for these methods, such as data hungriness, the risk of automation, increasing complexity of model building, the no free lunch idea, and the winner’s curse.