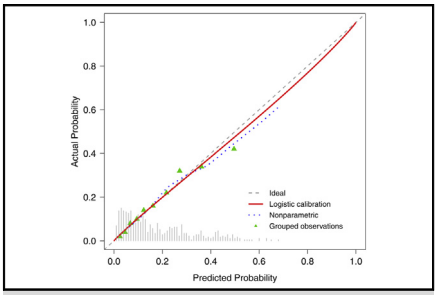
External model validation of binary clinical risk prediction models in cardiovascular and thoracic surgery

Graeme L. Hickey, PhD,[a] and Eugene H. Blackstone, MD[b]

ABSTRACT

Clinical risk-prediction models serve an important role in healthcare. They are used for clinical decision-making and measuring the performance of healthcare providers. To establish confidence in a model, external model validation is imperative. When designing such an external model validation study, thought must be given to patient selection, risk factor and outcome definitions, missing data, and the transparent reporting of the analysis. In addition, there are a number of statistical methods available for external model validation. Execution of a rigorous external validation study rests in proper study design, application of suitable statistical methods, and transparent reporting. (J Thorac Cardiovasc Surg 2016;152:351-5)



A calibration plot.

Central Message

External validation of binary clinical risk-prediction models is vital. We provide strategies for accomplishing this.

Perspective

The important role of clinical risk-prediction models for clinical decision-making and healthcare provider monitoring requires that they be externally validated. A model that has poor calibration or discrimination can result in misleading conclusions and suboptimal decision-making. This article highlights the key concepts.

See Editorial Commentary page 356.

Clinical risk-prediction models (CRPMs; also known as prognostic models or risk score models) serve an important role in healthcare,[1] particularly for binary adverse events (in-hospital, 30-day, or operative mortality) after cardiac, thoracic, and vascular surgery. These models may be applied to 3 different objectives: (1) to assess patient risk, which surgeons and patients can then factor in to healthcare decisions; (2) to stratify risk, both for clinical decision making and for determination of inclusion criteria in a controlled randomized trial[2]; and (3) to assess and compare healthcare outcomes among providers (benchmarking). The comparison of observed and expected outcomes, accounting for statistical uncertainty, can identify underperforming healthcare providers for quality improvement interventions.[3]

The wide-ranging importance of CRPMs in the cardiovascular specialty means that stakeholders must have confidence in them. A poorly performing model can lead to suboptimal decision making, misinformed patients, false reassurance of a healthcare provider's performance, or unfair stigmatization of a healthcare provider. Confidence is established by validating the model.[4]

Model validation can be internal, temporal, or external. Internal model validation is one element of CRPM development, usually published alongside the model to confirm that the model performs well for the training data. External validation, which evaluates the generalizability (or transportability) of the model to other groups of patients, is fundamental to demonstrating that a model is appropriate for adoption in clinical practice.[4] In cardiovascular and thoracic surgery, the majority of CRPMs encountered will

From the [a]Department of Biostatistics, University of Liverpool, Liverpool, United Kingdom and [b]Heart and Vascular Institute, Cleveland Clinic, Cleveland, Ohio.

Received for publication March 1, 2016; revisions received March 22, 2016; accepted for publication April 2, 2016; available ahead of print May 20, 2016.

Address for reprints: Graeme L. Hickey, PhD, Department of Biostatistics, University of Liverpool, Waterhouse Building (Block F), 1-5 Brownlow St, Liverpool L69 3GL, United Kingdom (E-mail: graeme.hickey@liverpool.ac.uk).

0022-5223/$36.00

http://dx.doi.org/10.1016/j.jtcvs.2016.04.023

Abbreviations and Acronyms
AUROC = area under the receiver operating characteristic curve
CRPM = clinical risk-prediction model
STS = Society of Thoracic Surgeons

predict binary outcomes, which were created using multivariable regression techniques, in particular logistic regression. Therefore, we focus our discussion here on this area. However, the general principles and need for external validation apply to other outcome types and models, such as time-to-event data,[5,6] as well as to nonregression techniques, such as machine learning approaches.[7]

MODEL PERFORMANCE CONCEPTS

Performance of CRPMs is typically assessed based on 2 important features: calibration and discrimination.[6] Calibration refers to the accuracy of the model for predicting events relative to observed events in groups of patients. For example, if the mean predicted event occurrence is 5% in a patient group but the observed event occurrence is 10%, then we conclude that the model is not well calibrated because it underpredicts.

Discrimination refers to the ability of a model to distinguish between patients who experienced the event and those who did not. Discrimination is measured using the area under the receiver operating characteristic curve (AUROC), also referred to as the concordance (c)-statistic or c-index.[5] This value has a meaningful interpretation. If we randomly select 2 patients, 1 patient who experienced the event and 1 who did not, then the AUROC is equivalent to the probability that the risk score attributed to the former is greater than that attributed to the latter. An AUROC of 1 indicates perfect classification; a value of 0.5 is equivalent to tossing a fair coin.

Other aspects of performance assessment include clinical usefulness, impact,[8] and overall performance measures such as the Brier score.[9]

DESIGNING AND REPORTING AN EXTERNAL VALIDATION

When designing a validation study, thought must be given to various key elements, including selection of patients, risk factor data, missing data, sample size, outcome definitions, study window size, and the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).

Selection of Patients

The selection of patients used to externally validate a CRPM might differ from those used to develop the model. These differences might be temporal or geographical, or related to clinical setting, inclusion or exclusion criteria, definitions, diagnostic techniques, or inherent baseline case mix differences between the 2 populations. It is important to highlight any differences that might affect model transportability between the validation sample and the original study sample, particularly with validation of general all-surgery models (eg, EuroSCORE) within procedural[10] or operative subgroups.[11]

Risk Factor Data

It goes without saying that calculating a risk score requires access to all variables that compose the risk score. One potential issue is conflict in variable definitions. For example, a registry that only collects binary data on whether pulmonary artery (PA) systolic pressure is >60 mmHg (a risk factor in the logistic EuroSCORE model) would not be able to compute the EuroSCORE II risk score, which includes model coefficients for PA systolic pressures of 31 to 55 mmHg and >55 mmHg. This is primarily an issue for retrospective validation studies, because clinical registries can be updated to capture contemporary risk score data.

Missing Data

One cannot calculate a risk score without access to data for variables that compose the CRPM. If a model contains a risk factor such as preoperative serum creatinine level but these data are sparsely available in the dataset, then in many cases the risk score cannot be calculated. Case-complete analyses—those that delete subjects with missing data for required variables—might lead to bias if those subjects are not representative of the whole population.[12] In certain cases, reasonable estimates and assumptions can be made based on clinical expertise or additional information in the dataset. A number of variables in Society of Thoracic Surgeons (STS) risk models have coefficients set to 0 for some variables in some models; if one is validating such a model, then missing data for such a variable is of no consequence. Alternatively, statistical imputation or subset analysis techniques might be applied to compensate.[13,14] If a validation study specifically excludes certain groups of patients (eg, emergency surgery, reoperations, or endocarditis), then imputation of 0 is an accurate and appropriate substitution, but the validation is only partial. In any case, it is always necessary to summarize the frequency of missing data and present methods for managing it and its assumptions.

Sample Size

Considerations regarding sample size should not be limited to randomized control trials. Single-center validation studies often will have a limited pool of subjects, especially for subgroup analyses, and increasing the sample size will require expanding the study period, which could come at a price (see the comment on calibration drift below). When designing a study, sample size (ie, number of subjects) alone is not enough; one also must consider effective