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)

Clinical risk-prediction models (CRPMs; also known as prognostic models or risk score models) serve an important role in healthcare, 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; 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.

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.

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. In cardiovascular and thoracic surgery, the majority of CRPMs encountered will
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, as well as to nonregression techniques, such as machine learning approaches.

MODEL PERFORMANCE CONCEPTS

Performance of CRPMs is typically assessed based on 2 important features: calibration and discrimination. 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. 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, and overall performance measures such as the Brier score.

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 or operative subgroups.

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. 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. 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

Abbreviations and Acronyms

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

August 2016

The Journal of Thoracic and Cardiovascular Surgery
sample size (ie, number of events). Relatively little attention has been given to this matter, but some studies have recommended a minimum of 100 events and 100 nonevents for validation studies, and in certain applications, larger effective sizes will be required to obtain adequate power.15,16

Outcome Definitions

Many well-known CRPMs in cardiac surgery, including the logistic EuroSCORE model17 and STS cardiac surgery risk models,18-20 predict early or operative mortality. Operative mortality is generally accepted to mean death within 30 days (or later if the patient has not been discharged within 30 days)21; however, other definitions of mortality exist, such as in-hospital mortality.22 Two large databases have reported operative mortality of 4.63% and 3.57%, compared with in-hospital mortality of 4.02% and 2.94%, respectively.23,24 In both cases, in-hospital mortality was approximately 0.6% lower. In-hospital mortality is generally easier to measure robustly, whereas 30-day mortality requires postdischarge follow-up for most patients.25 Therefore, it is common to see models validated against in-hospital mortality. In this example, we would expect the model to overpredict mortality relative to the observed data. It is reasonable to assess the model performance for this similar endpoint; however, this subtlety should be borne in mind when designing a study, particularly if the objective of the study is to compare models with different outcome definitions. Similar considerations apply to cases where the definition of a major postoperative complication used for model development differs from that in the validation dataset.

Study Window Size

One simple way to increase sample size in a validation study is to expand the study window. Validation of a CRPM over a substantially wide period can introduce a number of complexities, however. One potential issue is calibration drift.26,27 Multiple studies have demonstrated a decreasing ratio of observed mortality to mean logistic EuroSCORE with time. Changing risk profiles, other variables influencing mortality, and changes in the association of risk factors with outcomes can all contribute to this phenomenon. This situation prompted the introduction of the EuroSCORE II model22 and the series of contemporary STS models.18-20 Researchers should be aware of this, particularly when validating cardiac surgery CRPMs.

TRIPOD Statement

In recent years, reporting of biomedical research has been improved with guidelines such as the CONSORT statement28 for randomized trials and the PRISMA statement29 for systematic reviews and meta-analyses. Prompted by evidence of poor quality reporting in the CRPM literature, the recently published TRIPOD statement describes reporting guidelines for studies developing, validating, or updating a prediction model.30 We strongly encourage researchers to follow these guidelines and make use of the checklist for validating models. Examples of good practice and additional details have been published previously.31

METHODS FOR ASSESSING CALIBRATION

Hosmer–Lemeshow Test

The Hosmer–Lemeshow test is a frequently reported statistical test for assessing calibration in CRPMs. This test has several drawbacks, however.31-35 First, it is not easily interpreted: that is, it does not provide a measure of the magnitude of any miscalibration. Second, for slight deviations in calibration, the test is sensitive to sample size. Third, the classical version of the test is dependent on arbitrary groupings of patients. In some cases, the Hosmer–Lemeshow test remains a useful adjunct statistic, but should be included only as part of a more comprehensive assessment. Typically, the Hosmer–Lemeshow test refers to a test based on 10 groups composed by deciles of risk; however, one should be aware that there are variations of the test with regard to groupings (quantiles vs fixed cutpoints), number of groups (g), degrees of freedom of the χ^2 statistic (g-2 for internal vs g for external validation), and software implementation.35,36 Although g is typically selected to be 10, one must ensure that the cell counts are sufficient to justify the distributional approximation. Including a table of observed and expected events by binning group provides a useful summary, and allows for inspection of each term for fit, as recommended by Hosmer and Lemeshow (p. 188).37

Calibration Plot

If a standard Hosmer–Lemeshow test is performed, then a simple graph—the calibration plot—is a straightforward next step (Figure 1).4 Within each of the g groups, observed events are plotted against expected events. If the model is well calibrated, then these points should be close to the 45-degree line. The calibration plot can be augmented by overlaying a nonparametric smoothing curve (eg, loess) through the observed and predicted data or a calibration curve.4 In contrast to the Hosmer–Lemeshow test and basic calibration plot, these additional fits are not dependent on arbitrary groupings.

Calibration Curves

Cox’s calibration regression fits a logistic regression between the observed event and the log-odds transformed predicted values.40 A perfectly calibrated CRPM (deriving from a logistic regression model) yields an intercept of 0 and a slope of 1. These fitted regression models can be superimposed onto a calibration plot, giving an alternative graphical description of the miscalibration. Along with quantifying the degree of miscalibration, one also can simultaneously test whether the estimated parameters reject the null hypothesis of calibration. There are other related null hypotheses that can be tested for assessing calibration, as described by Steyerberg (p. 274).6

Other Tests

Although the Hosmer–Lemeshow test is ubiquitous in the biomedical CRPM literature, researchers can use a wide variety of statistical tests to assess model validation, such as the aforementioned calibration curve tests, the Spiegelhalter Z-test for calibration accuracy,40 and methods proposed by Stallard.31 Most of these can be calculated using routine software packages.38 There is no omnibus test of calibration; each approach has different merits and limitations. Therefore, it is important that researchers use a broad repertoire of methods to address the study questions.
CRPM validation study rests in proper study design, application of suitable statistical methods, and transparent reporting.

Conflict of Interest Statement
Authors have nothing to disclose with regard to commercial support.

References


Key Words: risk prediction model, external validation, calibration, discrimination, statistics

APPENDIX 1. R CODE TO PRODUCE FIGURE 1

```r
# If `rms` package not installed, run command
# install.packages(“rms”) library(rms)

## Simulate fake data:
## y = binary outcome
## x1, x2, x3 = covariates in the risk model
## n = sample size
set.seed(1)

n <- 1000 # 500 development + 500 validation
x1 <- runif(n) # covariate 1
x2 <- runif(n) # covariate 2
x3 <- runif(n) # covariate 3
logit <- -5 + 0.5*x1 + 2*x2 + 3.5*x3
P <- 1/(1 + exp(-logit))
y <- ifelse(runif(n) <= P, 1, 0) # outcomes
d <- data.frame(x1, x2, x3, y) # combined dataset

## Fit a risk prediction model to first half of the data
f <- lrm(y ~ x1 + x2 + x3, subset = 1:n)

## Use model to get predictions for second half of data
pred.logit <- predict(f, d[501:1000], )

## Validate prediction
val.prob(predat, y[501:1000], g = 10, riskdist = “predicted”)```