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

Graeme L. Hickey[1], Eugene H. Blackstone[2]

[1] University of Liverpool, Department of Biostatistics, Waterhouse Building (Block F), 1-5 Brownlow Street, Liverpool, L69 3GL, United Kingdom

[2] Cleveland Clinic, Heart and Vascular Institute Clinical Investigations, 9500 Euclid Avenue, JJ4, Cleveland, OH 44195, United States

The authors have no conflict of interest with material in this submission.

Words: ~ 2000

Corresponding Author:

Dr Graeme Hickey

University of Liverpool, Department of Biostatistics

Waterhouse Building (Block F), 1-5 Brownlow Street

Liverpool, L69 3GL, UK

Email: graeme.hickey@liverpool.ac.uk

Tel: +44 (0)151 794 9737

21 **CENTRAL PICTURE**

22

23 **[Use Figure (central).pdf]**

24

CENTRAL MESSAGE

[Characters + spaces = 117; Limit = 200]

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

INTRODUCTION

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 inclusion criteria in a controlled randomized trial,[2] and 3) to assess and compare healthcare outcomes among providers (benchmarking). The comparison of observed with 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 false stigmatization of the 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 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 a model is appropriate for adoption in clinical practice.[4] 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 to this area. However, the general principles and need for external

validation apply to other outcome types and models, e.g. time-to-event data,[5,6] as well as to non-regression techniques, e.g. machine learning approaches.[7]

MODEL PERFORMANCE CONCEPTS

Performance of CRPMs is typically based on assessing two important features: calibration and discrimination.[6]

Calibration is 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 the model is not well calibrated because it underpredicts.

Discrimination is 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 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] and concordance index, particularly for time-related events.

DESIGNING AND REPORTING AN EXTERNAL VALIDATION

When designing a validation study, thought must be given to several key elements.

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 two populations. It is important to highlight any differences that might affect model transportability between the validation and original study sample, particularly with validation of general all-surgery models (e.g. the 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 comprise 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 – 55 mmHg and >55 mmHg. This is primarily an issue for retrospective validation studies, as 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 comprise the CRPM. If a model contains a risk factor such as preoperative serum creatinine, 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. For example, 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, 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 (for example, emergency surgery, reoperations, or endocarditis), 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 will often have a limited pool of subjects, especially for subgroup analyses, and increasing the sample size will require widening the study period, which could come at a price (see comment on calibration drift below). When designing a study, sample size (number of subjects) alone is not enough; one must also consider effective sample size (number of events). Relatively little attention has been given to this matter, but some studies have recommended a minimum of 100 events and 100 non-events 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 predict early or operative mortality, including the logistic EuroSCORE[17] and STS Cardiac Surgery Risk Models.[18–20] 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 reported operative mortality to be 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 robustly measure, whereas 30-day mortality requires post-discharge 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 over-predict 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 that have 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.

Large study windows. One simple way to increase sample size in a validation study is to widen the study window. However, validation of a CRPM over a substantially wide period can introduce a number of complexities. One potential issue is calibration drift.[26,27] Multiple studies demonstrated that the ratio of observed mortality to mean logistic EuroSCORE was decreasing with time. Changing risk profiles, other variables influencing mortality, and changes in the association of risk factors with outcome can all contribute to this phenomenon. This prompted the introduction of the EuroSCORE II model[23] 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 statement[28] for randomized trials and the PRISMA statement[29] for systematic reviews and meta-analyses. Prompted by evidence of poor quality reporting in the CRPM literature, the recent 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 previously published.[31]

METHODS FOR ASSESSING CALIBRATION

Hosmer-Lemeshow test. The Hosmer-Lemeshow test is a frequently reported statistical test for assessing calibration in CRPMs. However, it has a number of drawbacks.[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 only be included 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, authors should be aware that there are variations on the test with regard to groupings (quantiles vs. fixed cut-points), number of groups (g), degrees of freedom of the chi-squared statistic (g-2 for internal vs. g for external validation), and software implementations.[35,36] While g is typically selected to be 10, one must ensure 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).[36]

Calibration plot. If a standard Hosmer-Lemeshow test is performed, then a simple graph—the calibration plot—is a straightforward next step (**Figure**).[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° line. The calibration plot can be augmented by overlaying a non-parametric smoothing curve (e.g. loess) through the observed and predicted data[37] or a calibration curve.[38] Contrary 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.[39] A perfectly calibrated CRPM (deriving from a logistic regression model) yields an intercept = 0 and a slope = 1. These fitted regression models can be superimposed onto a calibration plot, giving an alternative graphical description of the miscalibration. As well as quantifying the degree of miscalibration, one can also 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 also (p. 274).[6]

Other tests. The Hosmer-Lemeshow is ubiquitous in biomedical CRPM literature. However, researchers can take advantage of a wide variety of statistical tests to assess model validation, such as the aforementioned calibration curve test(s), the Spiegelhalter *Z*-test,[40] and methods proposed by Stallard.[41] Most can be calculated using routine software packages.[6,38] There is no omnibus test of calibration; each approach has different merits and limitations. Therefore, it is important that researchers employ a broad repertoire of methods to address the study questions.

MODEL UPDATING

A natural extension to the validation of a CRPM is the concept of updating an existing model. This might involve exploring whether a new biomarker improves a model (e.g. using net reclassification improvement measures[42]), recalibrating a model,[43] and, more recently, assessing whether multiple models can be combined to provide a more accurate prediction (e.g. meta-models and model averaging).[44] This expanding research area is especially important in an era of personalized medicine.[45]

CONCLUSIONS

External validation of CRPMs is necessary to demonstrate their predictive accuracy. Available models have likely been validated internally; however, using them in different settings, locations, and populations can result in relatively poor performance. CRPMs that have been overfitted during development will also often fail to generalise to the external validation sample. Calibration and discrimination must be measured in order to establish validity. There are multiple statistical approaches available to interrogate the calibration, with it being widely accepted that the ubiquitous Hosmer-Lemeshow test has limited utility. Execution of a rigorous CRPM validation study rests in proper study design, application of suitable statistical methods, and transparent reporting.

REFERENCES

1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338(b375):1317-1320.

2. Leon MB, Smith CR, Mack MJ, Miller DC, Moses JW, Svensson LG, et al. Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. *N Engl J Med*. 2010;363(17):1597-1607.

3. Bridgewater B, Hickey GL, Cooper G, Deanfield J, Roxburgh JC. Publishing cardiac surgery mortality rates: lessons for other specialties. *BMJ*. 2013;346:f1139.

4. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338(b605):1432-1435.

5. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361-387.

6. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer; 2009.

7. Nilsson J, Ohlsson M, Thulin L, Höglund P, Nashef SAM, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg*. 2006;132(1).

8. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338(b606):1487-1490.

9. Rufibach K. Use of Brier score to assess binary predictions. *J Clin Epidemiol*. 2010;63(8):938-939.

10. Grant SW, Hickey GL, Dimarakis I, Trivedi U, Bryan AJ, Treasure T, et al. How does EuroSCORE II perform in UK cardiac surgery; an analysis of 23 740 patients from the Society for Cardiothoracic Surgery in Great Britain and Ireland National Database. *Heart*. 2012;98(21):1568-1572.

11. Grant SW, Hickey GL, Dimarakis I, Cooper G, Jenkins DP, Uppal R, et al.

Performance of the EuroSCORE models in emergency cardiac surgery. *Circ Cardiovasc Qual Outcomes*. 2013;6(2):178-185.

12. Knol MJ, Janssen KJM, Donders RRT, Egberts ACG, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol*. 2010;63(7):728-736.

13. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol*. 2009;9:57.

14. Janssen KJM, Vergouwe Y, Donders a. RT, Harrell Jr FE, Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem*. 2009;55(5):994-1001.

15. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58(5):475-483.

16. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35(2):214-226.

17. Roques F. The logistic EuroSCORE. *Eur Heart J*. 2003;24(9):1-2.

18. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: Part 1 - coronary artery bypass grafting surgery. *Ann Thorac Surg*. 2009;88(1 Suppl):S2-S22.

19. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: Part 2 - isolated valve surgery. *Ann Thorac Surg*. 2009;88(1 Suppl):S23-S42.

20. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: Part 3 - valve plus coronary artery bypass grafting surgery. *Ann Thorac Surg*. 2009;88(1 Suppl):S43-S62.

21. Jacobs JP, Mavroudis C, Jacobs ML, Maruszewski B, Tchervenkov CI,

Lacour-Gayet FG, et al. What is operative mortality? Defining death in a surgical registry database: a report of the STS Congenital Database Taskforce and the Joint EACTS-STS Congenital Database Committee. *Ann Thorac Surg*. 2006;81(5):1937-1941.

22. Swinkels BM, Plokker HW. Evaluating operative mortality of cardiac surgery: first define operative mortality. *Netherlands Hear J*. 2010;18(7-8):344-345.

23. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. *Eur J Cardio-Thoracic Surg*. 2012;41:1-12.

24. Siregar S, Groenwold RHH, de Mol B a JM, Speekenbrink RGH, Versteegh MIM, Brandon Bravo Bruinsma GJ, et al. Evaluation of cardiac surgery mortality rates: 30-day mortality or longer follow-up? *Eur J Cardio-Thoracic Surg*. 2013;44(5):875-883.

25. Hickey GL, Grant SW, Cosgriff R, Dimarakis I, Pagano D, Kappetein AP, et al. Clinical registries: governance, management, analysis and applications. *Eur J Cardio-Thoracic Surg*. 2013;44(4):605-614.

26. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardio-Thoracic Surg*. 2013;43(6):1146-1152.

27. Hickey GL, Grant SW, Caiado C, Kendall S, Dunning J, Poullis M, et al. Dynamic prediction modelling approaches for cardiac surgery. *Circ Cardiovasc Qual Outcomes*. 2013;6:649-658.

28. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Trials*. 2010;11(32):1-8.

29. Moher D, Liberati A, Tetzlaff J, Altman DG, Grp P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):1-6.

30. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Circulation*. 2015;131(2):211-219.

31. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis Or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.

32. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35(9):2052-2056.

33. Marcin JP, Romano PS. Size matters to a model's fit. *Crit Care Med*. 2007;35(9):2212-2213.

34. Paul P, Pennell L, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Stat Med*. 2013;32:67-80.

35. Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *J Epidemiol Biostat*. 2000;5(4):251-253.

36. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. Second Edi. New Jersey: John Wiley & Sons, Inc.; 2000.

37. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517-535.

38. Harrell Jr FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer; 2001.

39. Cox DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45(3-4):562-565.

40. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Stat Med*. 1986;5(5):421-433.

41. Stallard N. Simple tests for the external validation of mortality prediction scores. *Stat Med*. 2009;28:377-388.

42. Pencina MJ, D'Agostino Snr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:157-172.

43. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23(16):2567-2586.

44. Debray TP, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models. *Stat Med*. 2014;33(14):2341-2362.

45. Su T-L, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res*. 2015:In press.

46. Harrell Jr FE. rms: Regression Modeling Strategies. 2015. http://cran.r-project.org/package=rms.

FIGURE LEGEND

Figure. A calibration plot for simulated data (n = 500). The green triangles denote the mean predicted and observed event probabilities for patients grouped into tenths using deciles. The grey dashed line denotes perfect calibration. A smoothing curve (blue dashed line) and the calibration curve (red solid line) are also overlaid. The distribution of calculated predicted probabilities is overlaid along the horizontal axis. A subset of various statistics useful for validating the model are also shown. This figure was generated using standard statistical software: the rms package for R (R Core Team, R Foundation for Statistical Computing, Vienna, Austria; version 3.1.2). Further details are given in Harrell (2001)[38] and Harrell (2015).[46] Code to reproduce this plot is given in the Appendix.

APPENDIX

R code to produce figure

```
# If 'rms' package not install, 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 = 1:500)

## Use model to get predictions for second half of data
pred.logit <- predict(f, d[501:1000, ])
phat <- 1 / (1 + exp(-pred.logit))

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