ARTICLE REVIEW

THE ROLE OF EXTERNAL VALIDATION STUDIES OF CLINICAL PREDICTIVE MODELS (CPMS) IN ACUTE RESPIRATORY DISTRESS SYNDROME (ARDS)

Hasna Afifah[1], Asri C Adisasmita[1], Fitriana Nur Rahmawati[2], Zulkifli Amin[2]
[1] Department of Epidemiology, Faculty of Public Health, Universitas Indonesia
[2] Division of Respirology and Critical Care Internal Medicine, Department of Internal Medicine, Faculty of Medicine Universitas Indonesia – Cipto Mangunkusumo Hospital

ABSTRAK

Model prediktif klinis atau sistem skoring saat ini makin populer dan mengakibatkan terlalu banyak model skoring yang ada namun studi yang melakukan validasi eksternal terhadap model-model tersebut masih sangat kurang. ARDS merupakan salah satu sindrom penyakit yang memiliki mortalitas dan morbiditas yang tinggi. Model skoring biasanya digunakan dalam memprediksikan luaran pada populasi yang memiliki risiko tinggi seperti pada ARDS. Pada telaah ini kami ingin memberikan gambaran tentang bagaimana studi eksternal harus dilakukan dan dilaporkan khususnya pada area ARDS. Pada area penelitian ARDS, sebagian besar studi validasi eksternal yang telah dilakukan memberikan laporan yang inadekuat, yaitu biasanya hanya menyebutkan diskriminasi saja dan tidak melaporkan kalibrasi. Kami merekomendasikan peneliti untuk mengikuti panduan TRIPOD yang merupakan panduan telaah kritis yang paling relevan dalam menilai dan melaporkan penelitian terkait model skoring. Studi validasi eksternal yang dilakukan dengan baik dan transparan dapat memudahkan klinisi dan peneliti lain dalam melakukan penilaian mengenai perfoma dan tingkat akurasi suatu model.

Kata kunci: acute respiratory distress syndrome, clinical predictive models, external validation, TRIPOD

ABSTRACT

Clinical Predictive Models (CPMs) have become increasingly popular in recent years and led to an overabundance of models while lacking validation studies. ARDS is a disease that still has a high mortality rate and burden. CPM has a role in predicting outcome in this high-risk population. We aim to provide a unifying overview of how an external study should be done and reported. In the field of ARDS research, external validation studies are hampered by inadequate assessment and reporting, mainly only mentioning discrimination and not calibration. TRIPOD guidance is the most reliable critical appraisal for CPMs. We suggest that TRIPOD guidance should follow CPMs to improve the methodology and analysis reports in external validation studies. Well-conducted and transparent external validation studies will make it easier for others to judge the performance of the predictive model.

Keywords: acute respiratory distress syndrome, clinical predictive models, external validation, TRIPOD

Correspondence :

Hasna Afifah

Department of Epidemiology, Faculty of Public Health, Universitas Indonesia

How to cite this article :

THE ROLE OF EXTERNAL VALIDATION STUDIES OF CLINICAL PREDICTIVE MODELS (CPMS) IN ACUTE RESPIRATORY DISTRESS SYNDROME (ARDS)

INTRODUCTION

Acute respiratory distress syndrome (ARDS) is a life-threatening condition characterized by acute, intense, and diffuse pulmonary inflammation causing complex damage to parenchyma or vasculature of the lungs.[1-3] The injury decreases the lung compliance and loses the permeability of pulmonary capillary endothelial and alveolar epithelial cells leading to refractory hypoxemia to usual oxygen therapy. Most studies report that the mortality rate is between 30-60%.[4-10] The mortality rate remains moderate to high in most developing countries. The newest large study conducted by Bellani et al. in 50 countries across five continents showed that the overall survival rate was 60.4% (95% CI = 58.7-62.2), and the hospital mortality from the study was approximately 34.9%, 40.3%, and 46.1% for those with mild, moderate, and severe ARDS respectively.[1]

The clinical predictive models (CPMs) play a role in predicting outcomes such as diagnosis and mortality. These CPMs were constructed from populations with various mortality rates and conditions. Consequently, when we apply CPMs on a new data set with a mortality rate and conditions different from the data set on which the model was constructed, the performance, especially the calibration value, can be altered. Inaccurate performance of CPMs will affect the prediction; it subsequently results in unnecessary or even harmful treatment. Different subpopulations, periods, outcome incidence/definitions, baseline characteristics, or diagnostic approaches across settings generally also affect the performance of CPMs. Before applying a CPM, it is essential to empirically evaluate its performance in the data set that was not used for the developed CPMs (external validation).

STUDIED CPMs IN ARDS

Clinicians usually adopt the widely used CPMs in Intensive Care Unit (ICU) settings, such as acute physiology and chronic health evaluation (APACHE) or simplified acute physiology score (SAPS) to predict ARDS outcomes. After the establishment of the new diagnostic criteria-Berlin definition in 2012, several CPMs for various ARDS subpopulations have also been proposed.[2-25] However, the abundance of CPMs and the lack of external validation studies of developed CPMs in the prognostic medical literature have led to research waste. It also obfuscates clinicians or healthcare providers in selecting the most useful CPMs. In the field of ARDS, inadequate study designs, sample size, lack of transparency, and incomplete reporting have become problems.[26] Conducting external validation should be a priority for assessing performance in other datasets including quantifying optimism from overfitting CPMs or poor statistical modelling during development, such as small sample sizes. Moreover, we can evaluate how good the transportability of CPMs is in a different setting. The more the external validation studies that show adequate performance, the more likely the CPMs will be useful.

Table 1. CPMs developed in ARDS population

Study	Objectives	Sample size	Events	Performance
Murray, et al (1988), (LIS)[24]	To identify patients with mild-moderate lung injury and severe acute lung injury (ARDS)	Not reported	Not mentioned	Not mentioned
Monchi, et al (1998)[25]	To predict mortality in ARDS	117 patients in developmental sample and 82 for validation	Mortality 65%	AUC 0,95; Hosmer-Lemeshow goodness-of-fit test p = 0,84
Cooke et al (2009)[26]	To predict mortality in acute lung injury	414 patients with non-traumatic ALI	28-day mortality	AUC 0,72; Hosmer-

		in the low tidal volume arm of trial, 459 for validation	26%	Lemeshow goodness-of-fit test p = 0.67
Gajic et al (2011), (LIPS)[27]	To identify patients at high risk of ALI early in the course of illness	5584 patients at risk	ALI 6,8%	AUC 0,8
Levitt et al (2013), EALI[28]	To identify patients with lung injury prior to requirement for positive pressure ventilation	256 patients, no validation	ALI requiring positive pressure ventilation 25%	AUC 0,85, Hosmer-Lemeshow p = 0,32
Lu S, et al (2013), (SESARDS)[12]	To predict mortality in ARDS patients	140 patients in developmental, 92 for validation	Mortality 64,2%	AUC 0,884, Hosmer-Lemeshow p = 0,382
Zhang et al (2015)[29]	To predict mortality in ARDS patients requiring mechanical ventilation	282 patients in developmental sample	Mortality 21,63%	AUC 0,85 95%CI: 0,79-0,9
Zhang et al (2015)[30]	To predict the use of corticosteroid in ARDS patient	745 patients	Mortality 27.52%	AUC 0.71, Hosmer-Lemeshow p = 0,7689
Go et al (2016)[31]	To estimate changes in the oxygenation index for 28-day mortality and VFD	1215 patients in developmental sample and 1185 for validation (ARMA, FACTT, ALVEOLI trial), another validation from ACURASYS trial	Mortality 28 day 24,5%, patients with fewer than 14 VFD 63,9%	Not mentioned
Villar et al (2016), (APPS)[32]	To predict in-hospital mortality of ARDS patients	300 patients in developmental sample and 300 for validation	Mortality 46.3%	AUC 0.755

KEY BASICS OF EXTERNAL VALIDATION

There are several important things that must be considered before conducting external validation. First, it should be clear whether the study has been done by an independent investigator (ideal) or by the author who developed the CPMs. Adequate sample size to externally validate the developed models using logistic regression is not well understood; therefore, it is recommended to use at least 100 events and ideally 250 (or more) to detect differences in performance of relevant models in an external validation study.[26,27] This suggestion is based on a hypothesis-testing framework (to detect pre-specified changes in performance) and simulation studies by Vergouwe, Y. et al.[28] and Collins, G, et al.[29] Many studies have adopted this guidance and described in transparent reporting of multivariable prediction models for individual prognosis or diagnosis (TRIPOD) guidance for a critical appraisal for model development and validation. A recent systematic review evaluating how well external validation studies were conducted showed that many external validation studies had fewer than 100 events. When we use a hundred samples with only a few events, there is a highly cautious interpretation because the performance results can be misleading.[27] Missing data often occur in both predictors and outcomes, including in development and validation studies. The handling of missing data should also be considered in the study report because it can lead to selection bias if handled inappropriately.[26] The outcome definition and diagnostic criteria may differ from how they were defined in the development of CPMs. For instance, in ARDS, diagnostic criteria also change over time. They have been improved from Murray's criteria to the American-European Consensus Conference (AECC) and now to the Berlin definition.[15,30,31] The Berlin definition is a more specific and generalized criterion and can be applied to a less heterogeneous population than the previous criteria. Different disease criteria, outcomes, or predictor definitions between validation and development studies can influence the accuracy of CPMs. Validation studies should refer to the original CPMs, and it is recommended to present a table summarizing the study characteristics, baseline characteristics, case-mix, and other critical elements between validation and development.[26, 27]

WHAT MAKES A GOOD CPM?

Calibration and discrimination are two key factors that should be reported in every validation study. These factors represent the general performance of a CPM. Other measurements of performance that can also be included are overall performance, reclassification, and clinical usefulness.[26, 27, 32]

Calibration is the degree of probability match between the predictions from the CPMs and the observed outcomes. It measures the accuracy of CPMs to predict the outcome of interest. Calibration performance receives little attention in the field of prognostic research and tends to be ignored if CPMs have good discrimination. This is a problem since poor calibration can lead to a misleading prediction. It has also been argued that calibration is the 'Achilles heel' of predictive analytics because poor calibration can make a CPM less useful than other CPMs with lower discrimination but well calibrated. When a CPM is developed in a dataset with low incidence, it may systematically show underestimated risk when used in a setting where the outcomes are high, and vice versa.[33-36] This problem is particularly crucial in the ARDS area since mortality rates vary between countries. An investigator should not focus on only one assessment since good discrimination does not guarantee good calibration and vice versa.[34]

Calibration may be well-calibrated in some ranges of predictive risk but not in others.[37] For example, an externally validated APACHE II can accurately estimate risk for ARDS patients in the middle range of score (40% mortality risk), but the CPM overestimates a higher score. Such poor calibration among patients with higher risk may or may not be a problem. It depends on the threshold used by clinicians for decision making. If the threshold is 40%, a CPM that overestimates by more than 40% would still be useful, and the overestimation in patients with higher risk would be irrelevant. The threshold may be subjective and different among clinicians, and it may need a good-calibration in almost all ranges of predicted risk. In such cases, it is necessary to update with recalibration or modification of CPM. How do we address this problem? That is the role of conducting an external validation study that is applied routinely in a clinical setting. External validation is the most substantial test of CPM. Published or a widely used CPM is not a guarantee that it has the same performance in other settings different from the one developed, even if it is applied to a population that is plausibly related or very similar to the development setting. We suggest that external validation be firstly conducted on CPMs that have been routinely used in our settings.

Calibration is preferably reported in several reporting measurements. We cannot conclude whether the CPM is well-calibrated or not if we only show one measurement. It is ideally reported as a calibration plot or graphically with the observed risk plotted on the y-axis and the predicted risk on the x-axis (Figure 1). This plot displays the magnitude of model miscalibration across the probability range. Calibration plots can be visualized using some statistical software, such as R or the pmcalplot module in STATA.[38]

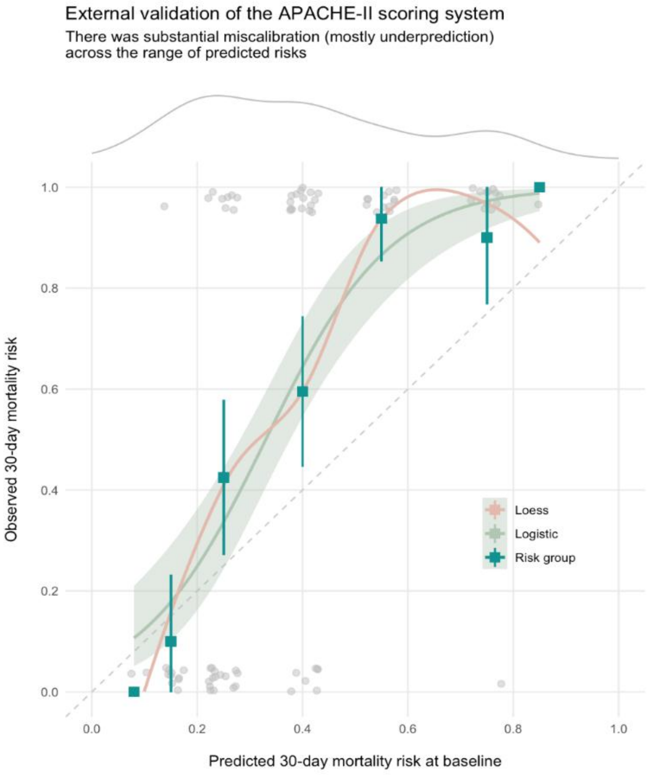

Figure 1. Calibration plot. An example of a calibration plot for risk prediction from a prognostic model with a binary outcome, produced using the R package.

In that plot, other metrics, such as intercept and slope, can be displayed. The CPM is considered well-calibrated if the slope is closer to 1, and the intercept is closer to 0. It means that the observed and predicted agreement is around the 45◌̊ line of the calibration plot or perfect calibration. More importantly, it reflects consistent calibration across a wide range of individuals.[39]

The intercept relates to calibration-in-the-large (α), which compares the mean of all predicted risks with the mean observed risks. This means that the prediction is systematically too low or too high. In binary outcomes, this can be measured by fitting logistic model for the probability of the outcome (P[Y = 1]) with the linear predictor (LPi) as a covariate (offset term).

$$\mathrm{logit}(P[Y = 1]) = \alpha + 1(\mathrm{LPi})$$

In model development, α = 0 and β or slope=1 for regression models. It means that the calibration-in-the-large should be close to zero for a well-calibrated model. In a validation study, in which the outcome of interest is different from when the CPM was developed, the value may deviate from zero ($\alpha < 0$ means systematic overprediction, while $\alpha > 0$ means systematic underprediction).[34]

The slope or mainly mentioned as calibration slope (β) is a measurement between the observed and predicted risk of the outcome across a range of predicted values.

$$\mathrm{logit}(P[Y = 1]) = \alpha + \beta(\mathrm{LPi})$$

When validating, the slope often deviates from 1 value. $\beta<1$ reflects an overfitting model (e.g., low probabilities are predicted too low and high probabilities are predicted too high). A value of $\beta<1$ can also be interpreted as a need for shrinkage of the regression coefficients in a CPM. $\beta > 1$ indicates that the predictions are too narrow.

A value β less than 1 is often found in external validation studies, consistent with the lack of adjustment for overfitting CPMs when they were developed. A value of slope β = 1 cannot reflect a good calibration without reporting the intercept or calibration plots. This can occur in the risk of a CPM being systematically overpredicted or underestimated. The slope β should be reported in conjunction with the intercept or calibration plot.[25, 39]

Many studies frequently use the Hosmer-Lemeshow goodness-of-fit test as a calibration test for logistic regression models. The test assesses the correspondence between predictions and observations by dividing the probability range (0 – 1) into n subgroups of the model population. It is based on arbitrarily dividing the data into risk strata and gives a p-value that is uninformative to the type of miscalibration, extends the miscalibration, and also suffers from low statistical power. Usually, it cannot provide sufficient penalties if the CPM is overfitting in the validation data. Consequently, it is recommended not to use this test for evaluating the calibration of CPM, and we should focus on reporting calibration as the calibration curve (graphic), intercept, and slope.[33, 34, 36]

Discrimination refers to the ability to distinguish between patients with a higher risk of having an outcome and those who will not. There are several ways to report discrimination; one of them is the c-statistic. In a binary outcome, the c-statistic is equivalent to the area under the ROC curve (AUC). It reflects the probability that the CPM scores or ranks from a randomly selected pair of patients with and without the outcome correctly ordered. A value of 1 indicates a perfect test. The value of 0.5 means the CPM cannot discriminate better than chance. This measurement does not reflect the prediction capability.[34-36] If a CPM can characterize a patient in the correct order, such a patient has a predictive risk of 2% having an outcome, and the other one who does not experience an outcome has a predictive risk of 2.1%. It always correctly ranks between such kind of pair, while it may have a miscalibration on the prediction value compared to their true or observed risk.

CONCLUSIONS

In conclusion, the investigator should present at least the suggested measurement of calibration and discrimination on the report. Decisions are often based on risk, so estimated risk by CPM should be reliable, and poor calibration can make a CPM useless and even harmful. Nevertheless, a perfect calibration is utopian; we aim for a CPM that can be clinically useful and harmless.[33, 40] The TRIPOD statement provides guidelines for researchers reporting studies that develop a new CPM or validate an existing one.[27] Better quality and transparent investigations will make a more impactful contribution to the field of prognostic research.

REFERENCES

1. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, Gattinoni L, van Haren F, Larsson A, McAuley DF, et al.; LUNG SAFE Investigators; ESICM Trials Group. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. JAMA 2016;315:788–800.
2. Jegal Y, Lee S, Lee KH, Oh YM, Shim TS, Lim CM, Lee S, et al. The clinical efficacy of GOCA scoring system in patients with acute respiratory distress syndrome. J Korean Med Sci. 2008;23: 383–9.
3. Lu S, Cai S, Ou C, Zhao H. Establishment and evaluation of a simplified evaluation system of acute respiratory distress syndrome. Yonsei Med J 2013;54:935-41.
4. Saleh A, Ahmed M, Sultan I, Abdel-lateif A. Comparison of the mortality prediction of different ICU scoring systems (APACHE II and III, SAPS II, and SOFA) in a single-center ICU

subpopulation with acute respiratory distress syndrome. Egyptian Journal of Chest Diseases and Tuberculosis 2015;64:843–8.

5. Lin CY, Kao KC, Tian YC, Jenq CC, Chang MY, Chen YC, et al. Outcome scoring systems for acute respiratory distress syndrome.Shock. 2010;34:352-7.
6. Sekulic AD, Trpkovic SV, Pavlovic AP, Marinkovic OM, Ilic AN. Scoring systems in assessing survival of critically ill icu patients. Med Sci Monit 2015;21:2621-9.
7. Kamal M, Khan AN, Ali G. A comparison of APACHE II and APACHE IV scoring systems in predicting outcome in patients with acute lung injury (ALI) and the adult respiratory distress syndrome (ARDS) in intensive care unit (ICU). RMJ. 2013;38: 234-8.
8. Atta MS, Mahrous AAA, Hassanien AA.Developing prevention model of acute lung injury: Validity of lung injury prediction score and risk panel. Egyptian Journal of Chest Diseases and Tuberculosis 2013;62:675–85.
9. Bassford CR. [internet]. 11β-hydroxysteroid dehydrogenase glucocorticoid metabolism within the lung and its influence on macrophage function in the acute respiratory distress syndrome. [cited 2016 june 23]. Available from: http://wrap.warwick.ac.uk/49584/1/wrap_thesis_bassford_2011.pdf
10. Villar J, Blanco J, Kacmarek RM. Acute respiratory distress syndrome definition: do we need a change?. Curr Opin Crit Care 2011;17(1):13-7.
11. Bauman ZM, Gassner MY, Coughlin MA, Mahan M, Watras J. Lung injury prediction score is useful in predicting acute respiratory distress syndrome and mortality in surgical critical care patients. Crit Care Res Pract [internet]. 2015 [cited 2016 June 20]; 2015:[about 8 p]. Available from: http://www.hindawi.com/journals/ccrp/2015/157408/
12. Cartin-Ceba R, Trillo-Alarez C, Kashyap R, Kolicic M, Dong Y, Poulose J, et al. Derivation of a lung injury prediction score (lips) to identify patients at high risk of ards at the time of hospital admission. Respir Crit Care Med 2009; A4653.
13. Kor DJ, Lingineni RK, Gajic O, et al. Predicting risk of postoperative lung injury in high-risk surgical patients: a multicenter cohort study. Anesthesiology. 2014;120:1168-81.
14. Turenne E, Carmelle M, Hou PC, Mitani A, Barry JM, Kao EY, et al. Lung injury prediction score for the emergency department: first step towards prevention in patients at risk. Int J Emerg Med 2012;5:33.
15. Murray JF, Matthay MA, Luce JM, et al. An expanded definition of the adult respiratory distress syndrome. Am Rev Respir Dis 1988;138:720-3
16. Monchi M, Bellenfant F, Cariou A, et al. Early predictive factors of survival in the acute respiratory distress syndrome. A multivariate analysis. Am J Respir Crit Care Med 1998;158:1076-81
17. Cooke CR, Kahn JM, Caldwell E, et al. Predictors of hospital mortality in a population-based cohort of patients with acute lung injury. Crit Care Med 2008;36:1412-2.
18. Gajic O, Dabbagh O, Park PK, et al. Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort study. Am J Respir Crit Care Med 2011;183:462-70.
19. Levitt JE, Calfee CS, Goldstein BA, et al. Early acute lung injury: criteria for identifying lung injury prior to the need for positive pressure ventilation*. Crit Care Med 2013;41:1929-37
20. Zhang Z, Chen L. The association between fluid balance and mortality in patients with ARDS was modified by serum potassium levels: a retrospective

study. PeerJ. 2015;3:e752. Published 2015 Feb 10. doi:10.7717/peerj.752

21. Zhang Z, Chen L, Ni H. The effectiveness of Corticosteroids on mortality in patients with acute respiratory distress syndrome or acute lung injury: a secondary analysis. Sci Rep. 2015;5:17654. Published 2015 Dec 2. doi:10.1038/srep17654
22. Go L, Budinger GR, Kwasny MJ, et al. Failure to Improve the Oxygenation Index Is a Useful Predictor of Therapy Failure in Acute Respiratory Distress Syndrome Clinical Trials. Crit Care Med 2016;44:e40-4
23. Villar J, Ambrós A, Soler JA, et al. Age, PaO2/FIO2, and Plateau Pressure Score: A Proposal for a Simple Outcome Score in Patients With the Acute Respiratory Distress Syndrome. Crit Care Med 2016;44:1361-9.
24. Novac M, Dragoescu A, Stanculescu A, Duca L, Cenea D. The predictive value of scores used in intensive care unit for burn patients prognostic. Curr Health Sci J 2014; 40:253-9.
25. Huang L, Li T, Xu L, Hu XM, Duan DW, Li ZB, et al. Performance of Multiple Risk Assessment Tools to Predict Mortality for Adult Respiratory Distress Syndrome with Extracorporeal Membrane Oxygenation Therapy: An External Validation Study Based on Chinese Single-center Data. Chin Med J (Engl). 2016;129(14):1688-95.
26. Collins, G.S., de Groot, J.A., Dutton, S. et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol **14,** 40 (2014). https://doi.org/10.1186/1471-2288-14-40
27. Moons KG, Altman DG, Reitsma JB, 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. doi:10.7326/M14-0698
28. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5):475-483. doi:10.1016/j.jclinepi.2004.06.017
29. 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: 214–226
30. Bernard GR, Artigas A, Brigham KL, et al. Report of the American-European consensus conference on ARDS: definitions, mechanisms, relevant outcomes and clinical trial coordination. The Consensus Committee. Intensive Care Med 1994;20:225-32
31. ARDS Definition Task Force , Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin Definition. JAMA 2012;307:2526-33
32. Chen L. Overview of clinical prediction models. Ann Transl Med. 2020 Feb;8(4):71. doi: 10.21037/atm.2019.11.121. PMID: 32175364; PMCID: PMC7049012
33. Van Calster, B., McLernon, D.J., van Smeden, M. et al. Calibration: the Achilles heel of predictive analytics. BMC Med **17,** 230 (2019). https://doi.org/10.1186/s12916-019-1466-7
34. Riley RD, van der Windt D, Croft P, Moons KG. (Eds.). (2019). Prognosis Research in Healthcare: Concepts, Methods, and Impact. Oxford University Press.
35. Steyerberg, E. W. (2009). Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer.
36. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

37. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA*. 2017;318(14):1377–1384. doi:10.1001/jama.2017.12126

38. Joie Ensor & Kym IE. Snell & Emma C. Martin, 2018. "PMCALPLOT: Stata module to produce calibration plot of prediction model performance," Statistical Software Components S458486, Boston College Department of Economics, revised 04 Jan 2020.

39. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure?. J Clin Epidemiol. 2020;118:93-99. doi:10.1016/j.jclinepi.2019.09.016

40. Calster, B.V. and Steyerberg, E.W. (2018). Calibration of Prognostic Risk Scores. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). doi:10.1002/9781118445112.stat08078