European Heart Journal: Acute Cardiovascular Care

Trade-off between discrimination and calibration in risk scores: a perspective from the SOFA

--Manuscript Draft--

Manuscript Number:	ACC-D-22-00146
Full Title:	Trade-off between discrimination and calibration in risk scores: a perspective from the SOFA
Article Type:	Invited Editorial/Commentary/Educational
Section/Category:	General Intensive Care
Keywords:	Risk stratification; risk score; in-hospital mortality
Corresponding Author:	Xavier Rossello, MD, PhD Madrid, Madrid SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Xavier Rossello, MD, PhD
First Author Secondary Information:	
Order of Authors:	Xavier Rossello, MD, PhD
Order of Authors Secondary Information:	
Abstract:	N/A

This is an postprint (Accepted Manuscript) of an article published by Oxford in European Heart Journal. Acute Cardiovascular Care on 7 June 2022,available online: https://doi.org/10.1093/ehjacc/zuac036

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Trade-off between discrimination and calibration in risk scores: a perspective from the SOFA

Running head: Discrimination and calibration in risk scores

Xavier Rossello[1,2,3,4]

[1]*Cardiology Department, Health Research Institute of the Balearic Islands (IdISBa), Hospital Universitari Son Espases, Palma, Spain*

[2]*Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain*

[3]*Facultad de Medicina, Universitat de les Illes Balears, Palma, Spain*

[4]*Medical Statistics Department, London School of Hygiene and Tropical Medicine, London, United Kingdom.*

Address for correspondence:

Xavier Rossello, Cardiology Department, Hospital Universitari Son Espases, Palma, Spain & Health Research Institute of the Balearic Islands (IdISBa), Palma, Spain.
Email: fjrossello@ssib.es

Keywords

Risk stratification; risk score; in-hospital mortality

EDITORIAL

This editorial refers to 'Predictive ability of the sequential organ failure assessment score for in-hospital mortality in patients with cardiac critical illnesses: a nationwide observational study', by Yuji Nishimoto *et al.* pp. XXX–XXX.[1]

Risk stratification dates to the first coronary care unit (CCU) around 55 years ago, when Killip and Kimball stratified the severity of heart failure (HF) in patients with acute myocardial infarction (AMI) using a bedside tool that is still widely used nowadays.[2,3] In parallel to the evolution from traditional CCUs to modern cardiac intensive care units (CICUs), cardiologists have moved from an AMI-centred management to a more complex clinical situations involving not only patients with AMI, but also with advanced HF, cardiogenic shock, or multi-systemic critical illness.[4] In the current context, risk prediction tools applied to patients with a single primary cardiovascular diagnosis (e.g., GRACE for AMI patients) are less useful than new risk stratification tools addressing the complexities of the CICU population. In this sense, the Sequential Organ Failure Assessment (SOFA) score is gaining interest among CICU professionals.

In this issue of *European Heart Journal – Acute Cardiovascular Care*, Yuji Nishimoto *et al.* [1] focused on the predictive ability of the SOFA score in surgical and non-surgical patients with cardiac critical illnesses. Using 52,819 patients from the Japanese nationwide Diagnosis Procedure Combination database, the authors defined patients with cardiac critical as those admitted to the intensive care unit (ICU) and primarily managed by cardiologists or cardiovascular surgeons. They evaluated the discriminatory ability of the SOFA score for in-hospital mortality (6.8% in this cohort) using the area under the receiver operating characteristic curve (ROC) by subgroups of patients who underwent cardiac surgeries (64% of them) and non-surgical patients. The ROCs were 0.76 (95% CI, 0.74–0.77) for the former subgroup, and 0.83 (95% CI, 0.83–0.84) for the latter subgroup.

This study draws clinically meaningful conclusions, and their main findings are readily applicable into clinical practice, though several methodological considerations are worth being discussed. Particularly, those related to the study predictors (mortality vs severity of illness risk), and those related to assessment of the risk score performance.

When it comes to predict risk in the CICU, a clear distinction should be made for mortality and severity of illness risk scores.[4] Although patients with higher severity of illness are inherently at higher risk of death, these two concepts do not have the same clinical meaning in spite of their conspicuous overlapping. Mortality risk is the combination of baseline patient risk (driven by patient-specific factors) and the risk conferred by the severity of illness. Including non-modifiable prognostic factors not directly reflecting the severity of illness (e.g., age), overshadows the specific risk of the underlying pathophysiology, and do not provide insights into any potential benefit from changes in clinical management. In contrast, information exclusively related to the severity of illness more often drives the need for an intervention. Severity of illness risk scores use predictors reflecting physiological derangements, which quantify the severity of acute illness and predict short-term mortality risk. Among them, the SOFA score is one of the most widely used. It was originally developed to *"describe quantitatively and as objectively as possible the degree of organ dysfunction/failure over time"* in critically ill patients with sepsis.[5] Of note, their definition implied that the SOFA score was useful to improve the understanding of the natural history of single and multiple organ failure, as well as to assess the effects of new therapies on the course of organ(s) dysfunction. Vincent and colleagues reported that the SOFA score was not designed to predict outcome but to describe a sequence of complications in the critically ill.[5] Nevertheless, in the last ~25 years the prognostic value of the SOFA score to predict outcomes has been successfully tested in a broad range of critically ill conditions beyond sepsis[4], and it has become one of the most popular risk score in intensive care units.[4,6] Yet, there was still a gap to cover. Several studies had reported a high predictive ability of the SOFA score for in-

hospital mortality in patients with cardiac critical illnesses, though their performance was not as good for calibration as for discrimination.[4] Predictiveness according to type of cardiac critical patients (surgical *vs.* non-surgical) was also unknown.

By focusing too much on assessing discrimination through the area under the ROC curve, the study of Nishimoto *et al.*[1] overlooks other relevant dimensions needed to adequately assess the performance of a risk score (calibration and external validation).[7] Whilst calibration assessment was essentially reported by the authors in the supplementary material of the article, little is said about the external validation of their findings. Of note, the authors did not derive the SOFA score, but only gave a new application to that popular risk scale. Additionally, some kind of external validation can be inferred from their work (although the SOFA score has been largely validated in other populations[6], Japanese cohorts are generally underrepresented in validation studies).[8] Whilst how well calibration and external validation were assessed might be a subject for a different debate, the point is that the assessment of the predictive ability of the SOFA score for inhospital mortality in patients with cardiac critical illnesses was almost entirely focused on one single estimate: the area under the ROC curve, also known as c-statistic.[9]

Discrimination, measured as the area under the ROC curve is not the probability that patients are classified correctly but the probability that for any randomly selected pair of subjects, one with and one without the outcome, the model assigns a higher probability to the patient with the outcome.[10] Hence, the area of the ROC curve measures how well the SOFA score can rank those who will die at hospital, and those who will not. Of note, the c-statistic is a rank-based method that does not take distributions into account.[9] If their ranks are the same, a difference between two individuals who are at very low risk of dying, 2% *vs.* 3%, contribute the same to the c-statistic as two individuals on a very different risk of dying, such as 20% *vs.* 30%. The distribution of the SOFA score in survivors and non-survivors (observed in Figure 1 of the paper[1] is not relevant to estimate the c-statistic, what it really matters is how

well the predictions line up in rank order.[11] The use of c-statistics have one key advantage: discrimination can be summarized in a single number, which easy to interpret for clinicians with little statistical background.[12,13] Nevertheless, it must be acknowledged that some statistically significant differences across c-statistics might be hard to translate into clinically meaningful results (a difference between a c-statistic of 0.76 and 0.83 for surgical and non-surgical patients has not a clear clinical meaning when it comes to evaluate one individual). In contrast, the assessment of calibration requires higher methodological skills (which is generally a disadvantage), but provides plots with differences between observed and predicted outcomes that have a better clinical interpretation about when to use, and when not to use, the risk score (e.g., there can be one side of the plot with poorer calibration, usually the high-risk side). Both discrimination and calibration are key aspects of predictive performance in risk scores, and both should be thoroughly evaluated. Unfortunately, there is a trade-off between discrimination and calibration, and models cannot be perfect in both.[14] Calibration is consistently underevaluated, or at least underreported, in many risk score papers. Now, the key question for many risk score papers (**Graphical abstract**): is performing well a risk model that usually provides higher probabilities for the outcome in those who are going to experience it, even though these predicted probabilities do not necessarily match the actual outcome rate?

Acknowledgments

None.

Sources of funding

None.

Conflict of interest statement

The author declares that there is no conflict of interest.

Data availability statement

There is no original data in this editorial.

References

1. Nishimoto Y, Ohbe H, Matsui H, et al. Predictive ability of the sequential organ failure assessment score for in-hospital mortality in patients with cardiac critical illnesses: a nationwide observational study. *Eur Heart journal Acute Cardiovasc care*. Epub ahead of print February 2022. DOI: 10.1093/ehjacc/zuac011.

2. Killip T, Kimball JT. Treatment of myocardial infarction in a coronary care unit. A two year experience with 250 patients. *Am J Cardiol* 1967; 20: 457–64.

3. Rossello X, Ramis-Barceló MF, Raposeiras-Roubín S. Acute myocardial infarction with high Killip class: do geographic differences matter? *Eur Heart journal Acute Cardiovasc care* 2021; 10: 513–515.

4. Jentzer JC, Rossello X. Past, present, and future of mortality risk scores in the contemporary cardiac intensive care unit. *Eur Heart journal Acute Cardiovasc care* 2021; 10: 940–946.

5. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive care medicine* 1996; 22: 707–710.

6. Vincent JL, de Mendonça A, Cantraine F, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on 'sepsis-related problems' of the European Society of Intensive Care Medicine. *Crit Care Med* 1998; 26: 1793–1800.

7. Rossello X, Dorresteijn JA, Janssen A, et al. Risk prediction tools in cardiovascular disease prevention: A report from the ESC Prevention of CVD Programme led by the European Association of Preventive Cardiology (EAPC) in collaboration with the Acute Cardiovascular Care Association (ACCA) and the As. *Eur Heart Journal Acute Cardiovasc Care* 2020; 9: 522–532.

8. Wussler D, Kozhuharov N, Sabti Z, et al. External validation of the meessi acute heart failure risk score. *Ann Intern Med* 2019; 170: 248–256.

9. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; 115: 928–35.

10. Rossello X, Bueno H, Gil V, et al. MEESSI-AHF risk score performance to predict multiple post-index event and post-discharge short-term outcomes. *Eur Heart Journal Acute Cardiovasc Care* 2021; 10: 142–152.

11. Bueno H, Rossello X, Bardají A. Has the Fourth Universal Definition of Myocardial Infarction led to better diagnosis and risk stratification? *European Heart Journal* 2021; 42: 2562–2564.

12. Rossello X, González-Del-Hoyo M. Survival analyses in cardiovascular research, part I: the essentials. *Rev Española Cardiol (Engl Ed)* 2022; 75: 67–76.

13. Rossello X, González-Del-Hoyo M. Survival analyses in cardiovascular research, part II: statistical methods in challenging situations. *Rev Esp Cardiol (Engl Ed)* 2022; 75: 77–85.

14. González-Del-Hoyo M, Rossello X. Challenges and promises of machine learning-based risk prediction modelling in cardiovascular disease. *Eur Heart Journal Acute Cardiovasc*

care 2021; 10: 866–868.

Graphical abstract

Comprehensive risk score performance assessment

Click here to access/download;Graphical Abstract;Graphical abstract - risk scores.tif