Benedetto, U., & Dimagli, A. (2020). Commentary: To underfit and to overfit the data. This is the dilemma. *Journal of Thoracic and Cardiovascular Surgery*. https://doi.org/10.1016/j.jtcvs.2019.12.079

Commentary: To underfit and to overfit the data. This is the dilemma.

Umberto Benedetto MD, PhD[a], Arnaldo Dimagli MD[a]

[a] Bristol Heart Institute, University of Bristol, United Kingdom

The authors have no conflicts of interest.

Corresponding Author:

Umberto Benedetto

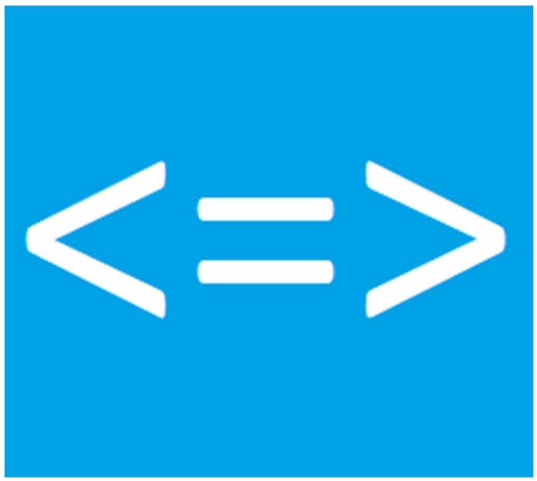
Office Room 84, Level 7,

Bristol Royal Infirmary, Upper Maudlin Street BS2 8HW, Bristol, United Kingdom

Tel. +44 (0) 117 3428854

email: umberto.benedetto@bristol.ac.uk

Word count: 495



Central message: Risk modelling should always consider the drawback of overfitting due to the inclusion of a large number of parameters.

Central picture: The right balance between overfitting and underfitting in risk modelling.

We read with great interest the paper from Mori et al [1] on the potential risk of poor performance of prediction models designed to be applied to heterogeneous group of surgical patients (so called universal model). They trained and tested a model including cholecystectomy, coronary artery bypass graft and esophagectomy. They concluded that the model performance was reduced when applied to specific subset of procedures, in particular with esophagectomy.

However, authors conclusions highlight possible limitation of these models and suggest that poor representation of low-volume case, model performance changes by the included case types, and variable effect sizes of unobserved covariates between case types can explain the poor performance observed in specific subset.

This manuscript looks at the oldest dilemma in risk modelling just from a different angle: the bias Variance Tradeoff [2]. In fact, a universal model will focus on a restricted number of variables which are common among different procedures. This model may be too simple and with very few parameters (underfitting) then it may have high bias (difference between the average prediction of our model and the correct value which we are trying to predict). On the other hand, if our model has large number of parameters to capture all possible aspects of individual procedures (overfitting), it will perform very well on training data but will have high error rates on test data (high variance).

The poor performance of the universal model tested by the authors can partially be attributed to the fact that variables were included in the model without a variable selection process such as (i.e. recursive feature elimination). We also should note that the poor performance of the universal model can be simply related to risk overestimation in subgroup at higher risk frequently observed with risk model based on logistic regression [3]. In fact, the universal model developed by the authors poorly performs in esophagectomy which is the procedure with the highest observed mortality. Moreover, authors have not specified the dataset time period and the poor performance can be partially explained by model calibration drift due to improvement in quality of care over the time or chance in case mix [4].

Broadly speaking, a universal model is very appealing because its implementation would allow the comparison of centres and surgeons' performance across a wide spectrum of surgical procedures. The statistical challenge remains and applies to any dataset: the balance between overfitting and underfitting the data.

References

1. Surgeons: Buyer Beware – Does ‘Universal’ Risk Prediction Model Apply to Patients Universally?
2. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc Natl Acad Sci U S A. 2019 Aug 6;116(32):15849-15854. doi: 10.1073/pnas.1903070116. Epub 2019 Jul 24.
3. Provenchère S, Chevalier A, Ghodbane W, Bouleti C, Montravers P, Longrois D, Iung B. Is the EuroSCORE II reliable to estimate operative mortality among octogenarians? PLoS One. 2017 Nov 16;12(11):e0187056.
4. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, Buchan I, Bridgewater B. 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 Cardiothorac Surg. 2013 Jun;43(6):1146-52.