



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>

Peer reviewed version

License (if available):
CC BY-NC-ND

Link to published version (if available):
[10.1016/j.jtcvs.2019.12.079](https://doi.org/10.1016/j.jtcvs.2019.12.079)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Elsevier at [https://www.jtcvs.org/article/S0022-5223\(20\)30066-0/fulltext](https://www.jtcvs.org/article/S0022-5223(20)30066-0/fulltext) . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

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

2 Umberto Benedetto MD, PhD^a, Arnaldo Dimagli MD^a

3 ^aBristol Heart Institute, University of Bristol, United Kingdom

4

5

6 **The authors have no conflicts of interest.**

7

8 Corresponding Author:

9 Umberto Benedetto

10 Office Room 84, Level 7,

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

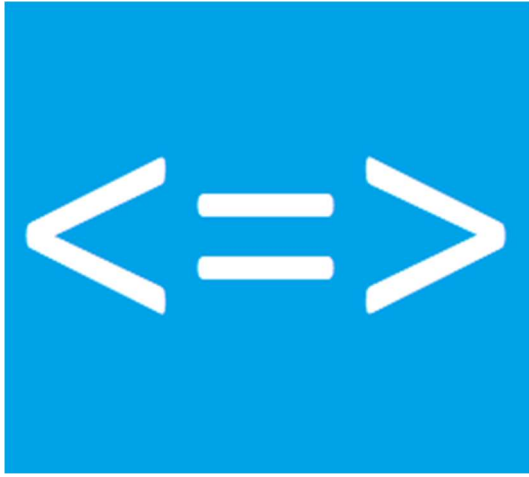
12 Tel. +44 (0) 117 3428854

13 email: umberto.benedetto@bristol.ac.uk

14

15

16 Word count: 495



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

3

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

5

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

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

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

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

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

14

15

16

17

18

19

20

21

22

1 References

- 2 1. Surgeons: Buyer Beware – Does ‘Universal’ Risk Prediction Model Apply to Patients
3 Universally?
- 4 2. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice
5 and the classical bias-variance trade-off. *Proc Natl Acad Sci U S A*. 2019 Aug
6 6;116(32):15849-15854. doi: 10.1073/pnas.1903070116. Epub 2019 Jul 24.
- 7 3. Provenchère S, Chevalier A, Ghodbane W, Bouleti C, Montravers P, Longrois D,
8 Lung B. Is the EuroSCORE II reliable to estimate operative mortality among
9 octogenarians? *PLoS One*. 2017 Nov 16;12(11):e0187056.
- 10 4. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, Buchan I,
11 Bridgewater B. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is
12 no longer suitable for contemporary cardiac surgery and implications for future risk
13 models. *Eur J Cardiothorac Surg*. 2013 Jun;43(6):1146-52.