Update to 'Deep-learning model for predicting 30-day postoperative mortality' (*Br J Anaesth* 2019; 123: 688–95)

Bradley A. Fritz*, Mohamed Abdelhack, Christopher R. King, Yixin Chen and Michael S. Avidan

St. Louis, MO, USA

*Corresponding author. E-mail: bafritz@wustl.edu

Keywords: anaesthesiology; death; deep learning; machine learning; postoperative complications

Editor—We write with an update regarding our recent article in the *British Journal of Anaesthesia*[1] describing a deep-learning model for intraoperative prediction of postoperative death. While performing additional work with the retrospective dataset we had used to train and test the described model, we discovered that many deceased patients had not had their records updated to reflect their death. We believe that all (or nearly all) the deaths reported in our previous work were true deaths (which we confirmed via manual chart review on a random subset of 50 patients). In the updated dataset, 2296 of 96 968 patients (2.4%) died within 30 days after surgery, including 1355 previously unlabelled deaths.

We repeated the model training, hyperparameter tuning, and model testing steps as described.[1] During hyperparameter tuning, our multipath convolutional neural network model performed better using 45-min epochs of intraoperative time series data than 60-min epochs. All the models, including our multipath convolutional neural network model and the comparison models, performed better in the updated dataset than in the original dataset. Area under the receiver operating characteristic curve and area under the precision-recall curve are shown in Table 1. As we described, an uninformative model will have area under the precision-recall curve equal to the incidence of the target, which was 0.024 in this population. The improvement in performance with the updated dataset was more marked for area under the precision-recall curve than for area under the receiver operating characteristic curve. As was true in our initial publication, the confidence intervals for the various models overlapped substantially for both performance metrics.

We also repeated the model calibration process using the histogram binning technique.[1] The observed incidence of mortality increased from 0.4% to 62% as the predicted probability of mortality increased, as shown in Fig. 1. The data points remain close to the diagonal, representing good calibration. The observed mortality in the highest-risk bin was much greater using the updated dataset (62%) than using the original dataset (18%), consistent with the model's overall better performance. The model was more likely to predict that patients would die if they had poor functional status or higher ASA physical status score; underwent emergency or cardiac surgery; or had coronary artery disease, congestive heart failure, deep venous thrombosis, diabetes mellitus, pulmonary hypertension, or chronic obstructive pulmonary disease. If the model predicted mortality, the prediction was more likely to be a true positive if the patient had poorer functional capacity or higher ASA physical status score, underwent emergency surgery, or had no history of hypertension or diabetes mellitus. If the model predicted survival, the prediction was more likely to be a false negative if the patient was older or male; had poorer functional capacity; underwent thoracic or vascular surgery; or had a cardiac comorbidity, chronic kidney disease, cancer, or anaemia.

In summary, the identification of additional patient deaths led to improved model performance by multiple metrics. The improved performance may be attributed partly to the more accurate labels for death in the updated dataset and partly to

Table 1 Performance of multipath convolutional neural network model (MPCNN) compared with deep neural network (DNN) without time series, random forest (RF), support vector machine (SVM), and logistic regression (LR). Both the long short-term memory (LSTM) and convolution neural network (CNN) methods of handling time-series data are presented. AUROC, area under receiver operating characteristic curve; AUPRC, area under precision-recall curve; CI, confidence interval

Model	Published results		Updated results	
	AUROC (95% CI)	AUPRC (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
MPCNN-LSTM	0.867 (0.835–0.899)	0.095 (0.085–0.109)	0.910 (0.897–0.924)	0.325 (0.280–0.372)
MPCNN-CNN	0.855 (0.822–0.887)	0.089 (0.077–0.100)	0.907 (0.894–0.920)	0.294 (0.251–0.339)
DNN	0.825 (0.790–0.856)	0.078 (0.068–0.088)	0.917 (0.905–0.930)	0.342 (0.296–0.389)
RF	0.848 (0.815–0.882)	0.078 (0.067–0.088)	0.923 (0.911–0.935)	0.409 (0.360–0.460)
SVM	0.836 (0.802–0.870)	0.072 (0.062–0.081)	0.913 (0.900–0.926)	0.314 (0.271–0.359)
LR	0.837 (0.803–0.871)	0.085 (0.074–0.096)	0.916 (0.904–0.929)	0.323 (0.279–0.368)

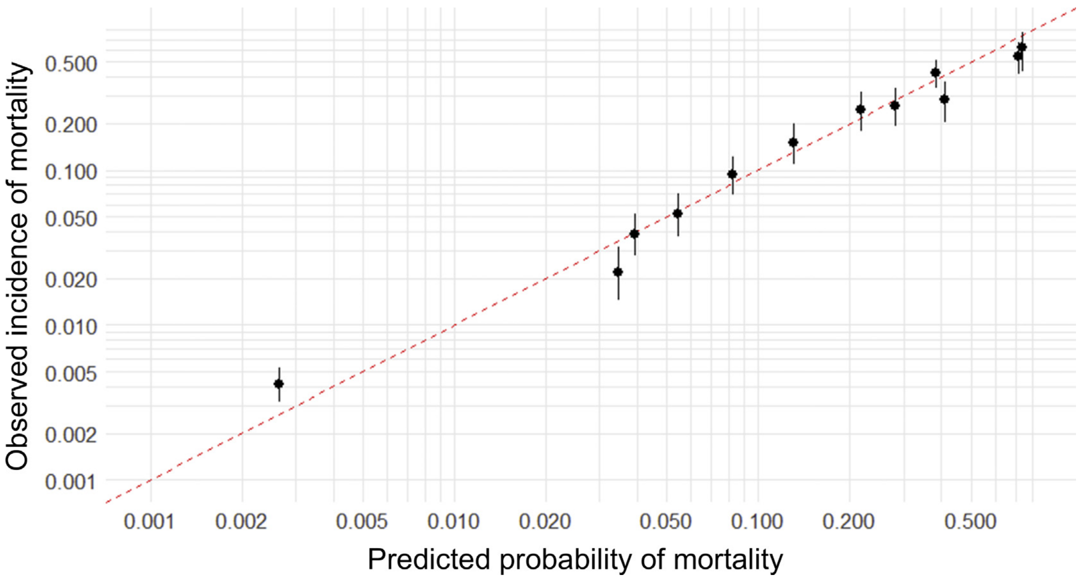

Fig. 1. Observed incidence of mortality *vs* calibrated predicted probability of mortality amongst patients in the test set (*n*=19 394). Predicted probabilities have been calibrated by applying the histogram binning technique in the validation set.

the increased incidence of death in the updated dataset. Our novel deep-learning model performed similarly to several comparison models, as indicated by the overlapping confidence intervals for area under the receiver operating characteristic curve and area under the precision-recall curve. The similarity in performance among the various models is qualitatively unchanged from the originally published results. It is easy for data quality issues to arise, particularly when large volumes of data are used or when datasets created for other reasons are repurposed for research use. Although the issue that we found did not lead to patient harm, similar issues in other clinical decision support situations may cause clinicians to take inappropriate actions that do lead to patient harm. Vigilance regarding data quality is a key step in machine learning, and this process does not stop once a model has been trained. Models intended for use in the clinical space must be continuously re-evaluated and updated. In our case, reusing a dataset for multiple analyses exposed a systematic error in outcome labels. A key takeaway from our experience is that incomplete labelling of the target variable can impair the performance of prediction models, even when robust analytic methods are applied.

Funding

US National Institute of General Medical Sciences (T32 GM108539) and the National Institute of Nursing Research (R01 NR017916).

Declarations of interest

MSA is an editor of the *British Journal of Anaesthesia*. The other authors have no conflicts to declare.

Reference

1. Fritz BA, Cui Z, Zhang M, et al. Deep-learning model for predicting 30-day postoperative mortality. *Br J Anaesth* 2019; **123**: 688–95

doi: 10.1016/j.bja.2020.03.024
Advance Access Publication Date: 7 May 2020

Nutritional factors in chronic musculoskeletal pain: unravelling the underlying mechanisms

Ömer Elma[1], Sevilay T. Yilmaz[1], Tom Deliens[1], Iris Coppieters[1,2], Peter Clarys[1], Jo Nijs[1] and Anneleen Malfliet[1,*]

[1]Brussels, Belgium and [2]Ghent, Belgium

*Corresponding author. E-mail: Anneleen.Malfliet@vub.be