Global Spine Journal
2022, Vol. 12(6) 1304–1305

DOI: 10.1177/21925682221085545
journals.sagepub.com/home/gsj

Letter to the Editor on "an Artificial Intelligence Approach to Predicting Unplanned Intubation Following Anterior Cervical Discectomy and Fusion" by Veeramani et al

Rafael De la Garza Ramos, MD[1,2] and Reza Yassari, MD[1,2]

To the Editor,

We read with great interest the study by Veeramani et al on artificial intelligence approaches to predict unplanned intubation after anterior cervical discectomy and fusion.[1] We agree that respiratory compromise can be devastating and that identification of high-risk patients is of paramount importance. We commend the authors for their efforts and wish to offer our insights. In our opinion, the main flaw of the presented work is that reporting accuracy, area under the curve (AUC), and Brier score to prove that these models are useful at predicting unplanned intubation is inappropriate and insufficient given the highly imbalanced classes, which in turn could be misleading to readers.

Veeramani et al[1] are dealing with an imbalanced classification problem with only 283 real positive cases (.51%) and 54,219 real negative cases (99.49%). Thus, the focus during model development should be towards the positive class, as models will be inherently biased towards patients who did not require reintubation. In such cases and what is unfortunately missing in the paper is data on recall (sensitivity) and precision (positive predictive value).[2]

Relying on accuracy as a measure for good performance is inappropriate in this dataset due to the accuracy paradox.[3-5] A model that guesses that zero patients will undergo unplanned intubation will have a 99.49% accuracy even if it misses all of the true positives. Obviously, such a model would not provide useful data and would be clinically unsound. Similarly, an AUC of .73 only tells us that this is the probability that a randomly chosen reintubated patient had a higher predicted risk compared to a randomly chosen non-reintubated patient. Note that this is *not* the probability that a patient requiring unplanned intubation is correctly classified (i.e., recall), or that a patient predicted to require reintubation will actually undergo reintubation (i.e., precision).[5] Lastly, the Brier score is also not very useful in this scenario. If the model estimates the unplanned intubation risk at .51% for a given patient and he/she does not require reintubation (the most likely outcome in this dataset), the Brier score will be almost perfect at .00026.

For these aforementioned reasons, not having data on the models' recall or precision is a significant limitation. Predicting rare events with an imbalanced dataset is challenging and advanced machine learning methods such as Synthetic Minority Over-Sampling Technique can be applied, but these are beyond the scope of this letter.[6] We encourage authors to construct confusion matrices for each of the models and plot the true positives, true negatives, false positives, and false negatives. Given that the *cost* of a false negative prediction can result in severe consequences for the patient, the models need to be optimized towards a higher recall value but without ignoring precision. We again commend the investigators for choosing to study such an important topic and hope these comments provide readers with additional perspective on the limitations of these algorithms before deciding to implement them in their practice.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

[1] Spine Research Group, Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, NY, USA
[2] Department of Neurological Surgery, Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, NY, USA

Corresponding Author:
Rafael De la Garza Ramos, Department of Neurological Surgery, Montefiore Medical Center, 3316 Rochambeau Avenue, 3rd floor, Bronx, NY 10467, USA.
Email: rafdelag@gmail.com

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Rafael De la Garza Ramos, MD https://orcid.org/0000-0002-5536-2514

References

1. Veeramani A, Zhang AS, Blackburn AZ, et al. An artificial intelligence approach to predicting unplanned intubation following anterior cervical discectomy and fusion. *Global Spine J*. 2022: 219256822110535. DOI:10.1177/21925682211053593.
2. Koehrsen W. *When Accuracy Isn't Enough, Use Precision and Recall to Evaluate Your Classification Model*. Chicago, IL: Accessed February 15, 2022 Builtin.comhttps://builtin.com/data-science/precision-and-recall.
3. Valverde-Albacete FJ, Peláez-Moreno C. 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PLoS One*. 2014;9:e84217. doi:10.1371/journal.pone.0084217
4. Lobo JM, Jiménez-Valverde A, RealAUC R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol Biogeogr*. 2008;17:145-151. DOI:10.1111/j.1466-8238.2007.00358.x
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115:928-935. DOI:10.1161/CIRCULATIONAHA.106.672402
6. Sowjanya AM, Mrudula O. Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Appl Nanosci*. 2022:1-12. DOI: 10.1007/s13204-021-02063-4