Artificial, but is it intelligent?

Michael R Levitt [1], Jan Vargas [2]

This editorial was not written by a chatbot, but it could have been.[1] The expansion of abilities in artificial intelligence and machine learning (AI/ML) has led to a dramatic uptake in a variety of disciplines, with particular excitement in medical diagnosis and prognosis. Aside from its increasingly common use in the detection of large vessel occlusion for rapid stroke triage,[2] recent applications of AI/ML in neurointervention have included patient selection[3] and prediction of functional outcomes in mechanical thrombectomy,[4–6] detection of catheter complications or undesirable embolization during endovascular intervention,[7–9] and identification of patients with procedurally challenging arterial anatomy,[10] among many others, employing AI/ML applications across large language models and computer vision.

The state of the science of AI/ML in clinical outcome prediction in particular was recently summarized in the pages of this journal.[11] A meta-analysis of 60 studies that used AI/ML to predict postoperative outcomes or complication after cerebrovascular or neuroendovascular surgery for stroke, aneurysm, or cerebral vascular malformation found relatively favorable performance compared with standard clinical prediction scales (area under the receiver operator characteristics curve (AUROC) >0.85 in most cases). Typically, such performance would be considered acceptable for clinical use. However, only 16.7% of such studies included external validation, and many had a high risk of bias.

Given the rapid evolution of AI/ML in neurointervention, it is tempting for the clinician to lean more and more on this technology for diagnosis, prognosis, and clinical decision-making. However, we identify areas of concern that must be addressed in future studies before widespread clinical adoption of this technology for such use. Common pitfalls can be categorized into two groups: poor study design, and a lack of proper statistical methodology in AI/ML. Given that study design is not unique to AI/ML, this work will focus on common mistakes with the latter, including data exploration, algorithm and metric selection; feature selection; and training and validation considerations.

DATA EXPLORATION, ALGORITHM AND METRIC SELECTION

One of the most common errors encountered in AI/ML is poor understanding of the underlying data. Descriptive statistics can (and should) be used to characterize subgroups and identify relationships between variables. For instance, regression models assume that independent variables are not correlated and are truly independent of each other. The basic assumptions of regression (linear and logistic) should be tested, such as ensuring a lack of multicollinearity among variables.

Many datasets in healthcare suffer from class imbalances, where the number of patients in one group is vastly overrepresented. For instance, attempting to model predictors of stroke in the general population is challenging, since stroke affects only a small percentage. Another example is the modeling of factors that contribute to chronic subdural hematoma (cSDH) recurrence following surgery, given that the recurrence rate of cSDH following surgery is up to 20%.[12] An imbalanced dataset coupled with poor metric and model selection can lead to the accuracy paradox: 'If the incidence of category A is found in 99% of cases, then predicting that every case is category A will have an accuracy of 99%.' Choosing appropriate metrics (such as precision and recall), utilizing under and over sampling techniques (for example, synthetic minority over-sampling technique (SMOTE)[13]), and selecting models that are resistant to class imbalance (such as tree based algorithms like XGBoost) are commonly used strategies to address this problem, though recent work on the effects of such corrections has generated controversy.[14 15]

In addition to the commonly used classification metrics of sensitivity, specificity, and AUROC, segmentation tasks (which are becoming increasingly popular in medical imaging) often use the Dice-Sørensen coefficient (also as the F1 score). Segmentation tasks rely on high-quality ground truth masks, which are often hand drawn by several human readers and cross-adjudicated. Critically, if the segmentation masks are small, then even minor mis-segmentations by a model can drastically lower the Dice score, affecting the perceived or actual performance of the AI/ML model.

FEATURE SELECTION

With datasets that contain many potential independent variables, selecting the 10 or 15 most important variables or features can often significantly boost model training. Most commonly this is done by utilizing least absolute shrinkage and selection operator (LASSO) regression, which shrinks the coefficients of unimportant features, or forward and backward stepwise feature selection in which each feature is either added or removed from a model, which is then validated using the Akaike information criterion. These approaches can help with model explainability and transparency, ensuring that features that are deemed important by the model make clinical sense. Tree-based models (for instance, XGBoost) inherently calculate the value of independent variables. Other models, however, such as neural networks or support vector machines, are not easily explainable. In these cases, Shapley analysis, which utilizes game theory methods to measure each variable's contribution to the final output of a model, can be employed.[16]

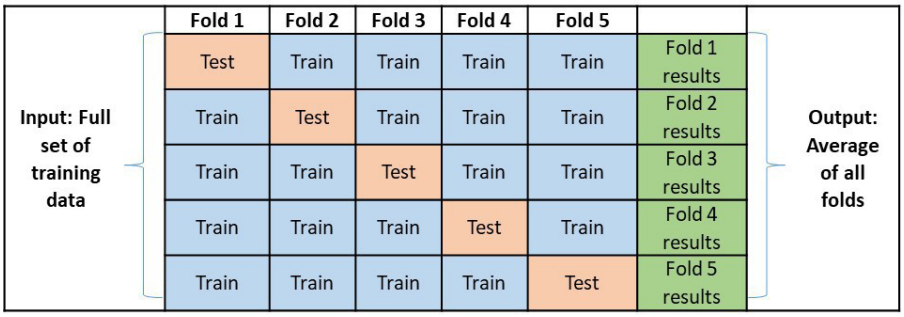

Figure 1 Schematic of cross-fold validation using five folds.

[1]Neurological Surgery, Radiology, Mechanical Engineering, Neurology, and Stroke & Applied Neuroscience Center, University of Washington, Seattle, Washington, USA
[2]Neurosurgery, Prisma Health Upstate, Greenville, South Carolina, USA

Correspondence to Dr Michael R Levitt; mlevitt@uw.edu

J NeuroIntervent Surg: first published as 10.1136/jnis-2024-022412 on 17 September 2024. Downloaded from http://jnis.bmj.com/ on October 9, 2024 by guest. Protected by copyright.

TRAINING AND VALIDATION

Common strategies for model training and validation include subdividing the dataset (such as patients or imaging studies) into training, testing, and validation subsets. The training set is used to construct the model, the testing set is used to adjust the model during training, and the validation set is used to measure the performance of the final model. When subdividing datasets in this way, there is a chance that classes and variables are not evenly distributed between sets, and so a model may overfit the training data and then underperform during testing and validation. Overfitting is a common problem in machine learning where a given model 'memorizes' the dataset and is no longer generalizable. A popular strategy for combating this is *k*-fold cross validation (figure 1), in which the training set is divided into *k* parts, uses *k*-1 parts for training and one for testing, and then repeats *k* times (or for *k* 'folds'), rotating the test set. Once this is done, the performance of all folds is averaged. Stratified *k*-fold cross validation ensures that variables are evenly distributed among parts.

Validation refers to the practice of running a trained model on a set of data that was previously isolated from the training data. Given the propensity of many AI/ML algorithms to overfitting, one of the most important commonly accepted practices is to validate models on external datasets (ideally, from other institutions or centers). This ensures model generalizability, which increases confidence in the performance abilities of the AI/ML algorithm.

CONCLUSION

The range of AI/ML-based clinical predictors for cerebrovascular and neuroendovascular procedural outcomes, and the use of these techniques in general in clinical medicine, may have the potential to greatly improve diagnosis and treatment. However, such research must be performed in a rigorous, generalizable, and reproduceable way. The neurointerventional field must keep in mind that such tools are only as good as the quality of the study design and data on which they are built, the majority of which are not suitable for clinical implementation.[17] As former IBM programmer George Fuechsel famously described in 1962, 'Garbage in, garbage out'.

X Michael R Levitt @DrMichaelLevitt and Jan Vargas @JanVargasMachaj

Contributors All authors contributed equally.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests MRL: Unrestricted educational grants from Medtronic and Stryker; consulting agreement with Medtronic, Aeaean Advisers, Metis Innovative and Genomadix; equity interest in Proprio, Stroke Diagnostics, Apertur, Stereotaxis, Fluid Biomed, and Hyperion Surgical; editorial board of Journal of NeuroInterventional Surgery, data safety monitoring board of Arsenal Medical. JV: Consulting agreement with Viz.AI, Imperative Care, Precision Neuro, Q'Apel, Medtronic and Microvention; equity interest in Viz. AI, Imperative Care, Borvo, Radical, Synchron; editorial board of Journal of NeuroInterventional Surgery.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; internally peer reviewed.

© Author(s) (or their employer(s)) 2024. No commercial re-use. See rights and permissions. Published by BMJ.

Check for updates

To cite Levitt MR, Vargas J. *J NeuroIntervent Surg* 2024;**16**:957–958.

Accepted 23 August 2024

J NeuroIntervent Surg 2024;**16**:957–958. doi:10.1136/jnis-2024-022412

ORCID iDs
Michael R Levitt http://orcid.org/0000-0003-3612-3347
Jan Vargas http://orcid.org/0000-0001-7164-1479

REFERENCES

1 Ray TR, Kellogg RT, Fargen KM, *et al*. The perils and promises of generative artificial intelligence in neurointerventional surgery. *J Neurointerv Surg* 2023;16:4–7.
2 Delora A, Hadjialiakbari C, Percenti E, *et al*. Viz LVO versus Rapid LVO in detection of large vessel occlusion on CT angiography for acute stroke. *J Neurointerv Surg* 2024;16:599–602.
3 Chen TC, Couldwell MW, Singer J, *et al*. Assessing the clinical reasoning of ChatGPT for mechanical thrombectomy in patients with stroke. *J Neurointerv Surg* 2024;16:253–60.
4 Pedro T, Sousa JM, Fonseca L, *et al*. Exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study. *J Neurointerv Surg* 2024.
5 Diprose JP, Diprose WK, Chien T-Y, *et al*. Deep learning on pre-procedural computed tomography and clinical data predicts outcome following stroke thrombectomy. *J Neurointerv Surg* 2024.
6 Liu C, Huang J, Kong W, *et al*. Development and validation of machine learning-based model for mortality prediction in patients with acute basilar artery occlusion receiving endovascular treatment: multicentric cohort analysis. *J Neurointerv Surg* 2023;16:53–60.
7 Sakakura Y, Masuo O, Fujimoto T, *et al*. Pioneering artificial intelligence-based real time assistance for intracranial liquid embolization in humans: an initial experience. *J Neurointerv Surg* 2024.
8 Masuo O, Sakakura Y, Tetsuo Y, *et al*. First-in-human, real-time artificial intelligence assisted cerebral aneurysm coiling: a preliminary experience. *J Neurointerv Surg* 2024.
9 Ghosh R, Wong K, Zhang YJ, *et al*. Automated catheter segmentation and tip detection in cerebral angiography with topology-aware geometric deep learning. *J Neurointerv Surg* 2024;16:290–5.
10 Canals P, Garcia-Tornel A, Requena M, *et al*. Deep learning-based model for difficult transfemoral access prediction compared with human assessment in stroke thrombectomy. *J Neurointerv Surg* 2024.
11 Hoffman H, Sims JJ, Inoa-Acosta V, *et al*. Machine learning for clinical outcome prediction in cerebrovascular and endovascular neurosurgery: systematic review and meta-analysis. *J Neurointerv Surg* 2024.
12 Levitt MR, Hirsch JA, Chen M. Middle meningeal artery embolization for chronic subdural hematoma: an effective treatment with a bright future. *J Neurointerv Surg* 2024;16:329–30.
13 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic Minority Over-sampling Technique. *J A I R* 2002;16:321–57.
14 van den Goorbergh R, van Smeden M, Timmerman D, *et al*. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc* 2022;29:1525–34.
15 Piccininni M, Wechsung M, Van Calster B, *et al*. Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models. *J Biomed Inform* 2024;155:104666.
16 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proc Neural Inf Proc Syst* 2017;4768–77.
17 Hager P, Jungmann F, Holland R, *et al*. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *N Med* 2024.

J NeuroIntervent Surg: first published as 10.1136/jnis-2024-022412 on 17 September 2024. Downloaded from http://jnis.bmj.com/ on October 9, 2024 by guest. Protected by copyright.