Reply: Developing Machine Learning Algorithms to Support Patient-centered, Value-based Carpal Tunnel Decompression Surgery

Luke Geoghegan, MBBS, BSc, MRCS*; Jeremy N. Rodrigues, FRCS, PhD*†; Chris J. Sidey-Gibbons, PhD‡; Paul H. C. Stirling, FRCS§; Jane E. McEachan, FRCS¶; Conrad J. Harrison, MRCS, PhD‖

Sir,

We thank Mantelakis and Khajuria[1] for opening the opportunity to explain some of the more technical elements of our recent article.[2]

Black box algorithms are indeed potentially dangerous, but our algorithms were not black boxes. One of the strengths of our article was the use of Shapley additive explanations (SHAP) to estimate the contribution of each feature to each decision. Contrary to what they claim, reading our article and supplementary material confirms that the variables they are concerned about made very little contribution to the algorithms' decisions.

A more useful discussion might have focused on the limitations of SHAP and similar procedures,[3] approaches to variable selection under traditional statistical and machine learning paradigms, or how model parsimony may have improved generalizability.

Mantelakis and Khajuria's understanding of data splitting, overfitting, and hyperparameter tuning is not entirely accurate. Through k-fold cross validation, the first dataset is iteratively partitioned into what they term "training" data (k-1-folds), and what they term "validation" data (1-fold, more commonly referred to as "test" data). Once hyperparameters are selected, parameters are fitted to the whole of the first dataset (k-folds). The second dataset acts as what Matelakis and Khajuria term a "test" set (usually known as the internal validation dataset).

A more valid criticism would be our use of internal validation only, rather than evaluating the models in an external sample. This is likely to overestimate performance, which we highlight as a limitation of our work. It would have been more insightful to challenge our use of data splitting in general, as this technique does not make the most of all available data. Other techniques, such as bootstrapping and adjusting for model optimism have been proposed, and these may have been preferable.[4]

Finally, the correspondence authors have confused area under the receiver operating characteristic curve with classification accuracy.

Our article certainly has limitations, and could be challenged by experts. For example, how appropriate is it to dichotomize patients into improved/not improved, based on distributionally estimated minimal important change statistics? How appropriate is it to dichotomize at all, and would a regression algorithm have been more appropriate than a classifier? Is class imbalance truly a problem? Our article lacks a calibration plot, or an assessment of net benefit and we could have perhaps shown closer adherence to the TRIPOD[5] and PROBAST guidance.[6]

Constructive criticism has an important role to play in improving publishing standards, and we fully support it. It is hoped that we have taken this opportunity to provide helpful feedback. We welcome further criticism of the article, and look forward to seeing original research contribution from Mantelakis and Khajuria to this field in the future.

Luke Geoghegan, MBBS, BSc, MRCS
Department of Plastic Surgery, Stoke Mandeville Hospital
Buckinghamshire Healthcare NHS Trust
Aylesbury, Buckinghamshire, UK
E-mail: lg1813@ic.ac.uk

*From the *Department of Plastic Surgery, Stoke Mandeville Hospital, Buckinghamshire Healthcare NHS Trust, Aylesbury, Buckinghamshire, UK; †Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, West Midlands, UK; ‡MD Anderson Center for INSPiRED Cancer Care, the University of Texas, Houston, Tex.; §Department of Trauma and Orthopaedic Surgery, Royal Infirmary of Edinburgh, Edinburgh, UK; ¶Department of Trauma and Orthopaedic Surgery, NHS Fife, Fife, UK; and ‖Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Headington, Oxford, UK.*

Plast Reconstr Surg Glob Open 2023; 11:e4744; doi: 10.1097/GOX.0000000000004744; Published online 25 January 2023.

REFERENCES

1. Mantelakis A, Khajuria A. Developing machine learning algorithms to support patient-centered, value-based carpal tunnel decompression surgery. *Plast Reconstr Surg Glob Open.* 2022;10:e4494.

Disclosure: *Dr. Harrison was funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684). Dr. Rodrigues was funded by an NIHR postdoctoral fellowship (PDF-2017-10-075). The other authors have no financial interest to declare. This document presents independent research funded by the NIHR. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.*

2. Harrison CJ, Geoghegan L, Sidey-Gibbons CJ, et al. Developing machine learning algorithms to support patient-centered, value-based carpal tunnel decompression surgery. Plast *Reconstr Surg Glob Open*. 2022;10:e4279.
3. Kumar IE, Venkatasubramanian S, Scheidegger C, et al. Problems with Shapley-value-based explanations as feature importance measures. In: *International Conference on Machine Learning*. Philadelphia, PA: MRS Press; 2020.
4. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
5. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg*. 2015;102:148–158.
6. Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51–58.