Neurosurgical Forum

LETTERS TO THE EDITOR

Indolent clinical and radiological effects of cervical disc arthroplasty

TO THE EDITOR: We read with interest the article by Yang et al.¹ (Yang MMH, Ryu WHA, Casha S, et al. Heterotopic ossification and radiographic adjacent-segment disease after cervical disc arthroplasty. *J Neurosurg Spine*. 2019;31(5):660–669). By retrospective analysis of 58 levels of cervical disc arthroplasty (CDA) in 45 patients with 1- or 2-level of CDA, the authors concluded that high-grade heterotopic ossification (HO) was predicted by residual exposed endplate (REE) and that CDA was protective against further degeneration at the index level. Although the authors should be applauded for their attempt to mitigate HO after CDA, we have to respectfully point out a few issues with the report.

The statistics could use some improvement. Although a total of 58 levels of CDA was stated and used for calculation in the report, there were only 57 levels demonstrated in Table 1 (33 patients who had 1-level and 12 patients who had 2-level operations; $33 + 2 \times 12 = 57$). Furthermore, for REE, the most critical results of the study should be divided into four groups (0–1, 1–2, 2–3, and > 3 mm) independently, rather than three groups (> 1 , > 2 , and > 3 mm) and additively calculated (Table 2). The omitted 0–1-mm REE very likely skewed the results of the study, given the already small sample size. Last, the separation of blue and red lines in Fig. 2 was repetitive and redundant because the high- versus low-grade HO was merely reciprocal.

The study also might be limited by neglect of many other reported risk factors of HO after CDA.^{2–8} For example, Wu et al. demonstrated that pre-existing calcified disc or spondylosis at the indexed level of CDA was associated with more than 9 times the incidence of HO.⁸ Smoking and the change in mobility after surgery also played an important role on the incidences of HO.^{4,9} Nevertheless, these cases of HO were usually indolent and rarely caused adverse radiological or clinical outcomes, or required reoperations.^{4–6,8}

The authors should be commended for demonstration of HO in various (5) brands of devices included in the study. The common phenomenon of the undesired bone formation—which fortunately seldom caused problems—after CDA prompts future investigations of the nature and prevention of HO. Given the proximity (at the index level) and demographic influence (patients' preoperative char-

acteristics) on the development of HO, it remained elusive whether CDA could hold or decelerate degeneration even at the index level. The protective effect of CDA on either the index or adjacent levels needs further evaluations.¹⁰

Wen-Cheng Huang, MD, PhD^{1,2}Yu-Chun Chen, MD, MSc^{2–4}Tsung-Hsi Tu, MD^{1,2}Hsuan-Kan Chang, MD^{1,2}Chao-Hung Kuo, MD^{1,2,5}¹Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan²School of Medicine, National Yang-Ming University, Taipei, Taiwan³Taipei Veterans General Hospital, Taipei, Taiwan⁴Institute of Hospital and Health Care Administration, National Yang-Ming University, Taipei, Taiwan⁵School of Biomedical Science and Engineering, National Yang-Ming University, Taipei, Taiwan

References

1. Yang MMH, Ryu WHA, Casha S, et al. Heterotopic ossification and radiographic adjacent-segment disease after cervical disc arthroplasty. *J Neurosurg Spine*. 2019;31(5):660–669.
2. Chang HK, Huang WC, Wu JC, et al. Should cervical disc arthroplasty be done on patients with increased intramedullary signal intensity on magnetic resonance imaging? *World Neurosurg*. 2016;89:489–496.
3. Chang PY, Chang HK, Wu JC, et al. Differences between C3–4 and other subaxial levels of cervical disc arthroplasty: more heterotopic ossification at the 5-year follow-up. *J Neurosurg Spine*. 2016;24(5):752–759.
4. Tu TH, Kuo CH, Huang WC, et al. Effects of smoking on cervical disc arthroplasty. *J Neurosurg Spine*. 2019;30(2):168–174.
5. Tu TH, Lee CY, Kuo CH, et al. Cervical disc arthroplasty for less-mobile discs. *J Neurosurg Spine*. 2019;31(3):310–316.
6. Tu TH, Wu JC, Huang WC, et al. The effects of carpentry on heterotopic ossification and mobility in cervical arthroplasty: determination by computed tomography with a minimum 2-year follow-up. *J Neurosurg Spine*. 2012;16(6):601–609.
7. Wu JC, Huang WC, Tsai HW, et al. Differences between 1- and 2-level cervical arthroplasty: more heterotopic ossification in 2-level disc replacement: Clinical article. *J Neurosurg Spine*. 2012;16(6):594–600.
8. Wu JC, Huang WC, Tu TH, et al. Differences between soft-disc herniation and spondylosis in cervical arthroplasty: CT-documented heterotopic ossification with minimum 2 years of follow-up. *J Neurosurg Spine*. 2012;16(2):163–171.
9. Chang HK, Chang CC, Tu TH, et al. Can segmental mobility be increased by cervical arthroplasty? *Neurosurg Focus*. 2017;42(2):E3.

10. Chang CC, Huang WC, Wu JC, Mummaneni PV. The option of motion preservation in cervical spondylosis: cervical disc arthroplasty update. *Neurospine*. 2018;15(4):296–305.

Disclosures

The authors report no conflict of interest.

Correspondence

Chao-Hung Kuo: chaohungk@gmail.com.

INCLUDE WHEN CITING

Published online January 31, 2020; DOI: 10.3171/2019.11.SPINE191369.

Response

We would like to thank Huang and colleagues for their interest in our study. In this study, we hypothesized increased REE to be a risk factor for the development of high-grade HO following CDA. Dr. Huang has appropriately suggested analyzing four independent REE groups to be more intuitive and appropriate. However, we did not have a sufficient sample size in each group to make this a strong argument. We divided the REEs into three groups (> 1 mm, > 2 mm, and > 3 mm) and found the > 2 -mm group to be significantly associated with the development of high-grade HO. We repeated our logistic regression analysis of REE on high-grade HO for all operated levels (i.e., not dividing the analysis into groups) and demonstrated that REE remained significantly associated with high-grade HO (OR 1.6 [95% CI 1.1–2.5], $p = 0.03$). From a statistical perspective, this allowed us to maximize the number of observations, minimize the variance, and maximize the power of the analysis. We believe this additional analysis, plus the results in our original paper, further supports the view that our conclusions are robust.

Similarly, our small sample size prevented adjustment for other known risk factors such as preoperative calcified disc or spondylosis at the index level of operation. With a sample size of 45 patients, we limited the number of adjustments made in our multivariable model to reduce the risk of obtaining biased regression coefficients and paradoxical associations.¹ With a larger sample size Dr. Huang's proposed analyses would be very interesting indeed. We thank Dr. Huang and colleagues for their insightful comments and hope they and others will expand on the results we have reported.

Michael M. H. Yang, MD, MSc, MBiotech
University of Calgary, AB, Canada

R. John Hurlbert, MD, PhD
Banner University Spine Program, University of Arizona, Tucson, AZ

References

1. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373–1379.

INCLUDE WHEN CITING

Published online January 31, 2020; DOI: 10.3171/2019.12.SPINE191376.

©AANS 2020, except where prohibited by US copyright law

Importance of calibration assessment in machine learning–based predictive analytics

TO THE EDITOR: With great interest, we recently read the article by Hopkins et al.,¹ in which the authors report on a predictive model for 30-day readmission after posterior lumbar fusion (Hopkins BS, Yamaguchi JT, Garcia R, et al. Using machine learning to predict 30-day readmissions after posterior lumbar fusion: an NSQIP study involving 23,264 patients. *J Neurosurg Spine*. 2020; 32(3):399–406). They had access to 23,264 patients for training of their model—a deep neural network. Among these patients, 1199 (5.15%) were readmitted. The authors internally evaluated their model in 20-fold cross-validation. They calculated a mean area under the curve (AUC) of 0.812 with sensitivity of 35.5% and a mean specificity of 99.5%. We highly commend the authors on their work in applying machine learning to this clinically important question. However, the authors do not assess the calibration of their predictive model.

When evaluating machine learning models for diagnosis or prediction of binary outcomes, two dimensions of performance need to be considered: First is discrimination—a model's ability to make correct binary prediction—which is commonly assessed using AUC, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and the F1 score. Second is calibration—the degree to which a model's predicted probability (ranging from 0% to 100%) corresponds to the actually observed incidence of the binary endpoint, which is commonly assessed using calibration curves, calibration slope and intercept, Brier score, the expected/observed ratio, and the Hosmer-Lemeshow test.²

While discrimination is practically always reported, many publications do not report calibration. While high discrimination measures and good calibration often coincide, such as is likely the case in the abovementioned publication, excellent discrimination does not necessarily mean that calibration is employable.³ Deep neural networks are especially prone to poor calibration, often massively skewing predicted probabilities toward “extreme” values of 1% and 99%.

Especially in clinical practice, calibration is crucial. For clinicians and patients alike, a predicted probability (e.g., being able to say to the patient, “Your risk is 7%”) is much more valuable than a binary yes/no prediction. Good calibration can often be attained through application of machine learning models with appropriate complexity in relation to the classification task at hand, such as logistic regression or generalized additive models. If poor calibration is observed and the pattern of miscalibration is consistent during resampling, recalibration techniques such as Platt scaling or isotonic regression can be applied.⁴ Lastly, models can also be primarily trained for measures of calibration, and intercepts can be adjusted.⁵

In conclusion, it is critically important to assess the calibration of clinical prediction models. A minimum of

calibration curve, slope, and intercept should be reported for every published model.

Victor E. Staartjes, BMed

Machine Intelligence in Clinical Neuroscience (MICN) Laboratory,
University Hospital Zurich, Clinical Neuroscience Centre, University of
Zurich, Switzerland

Julius M. Kernbach, MD

RWTH Aachen University, Aachen, Germany

References

1. Hopkins BS, Yamaguchi JT, Garcia R, et al. Using machine learning to predict 30-day readmissions after posterior lumbar fusion: an NSQIP study involving 23,264 patients. *J Neurosurg Spine*. 2020;32(3):399–406.
2. Debray TPA, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68:279–289.
3. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. <http://arxiv.org/abs/1706.04599>. Accessed January 30, 2020.
4. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*. New York, NY: ACM; 2005:625–632. <http://doi.acm.org/10.1145/1102351.1102430>. Accessed January 30, 2020.
5. Janssen KJM, Moons KGM, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61:76–86.

Disclosures

The authors report no conflict of interest.

Correspondence

Victor E. Staartjes: victoregon.staartjes@usz.ch.

INCLUDE WHEN CITING

Published online February 21, 2020; DOI: 10.3171/2019.12.SPINE191503.

Response

We greatly appreciate the feedback and constructive remarks in order to better our study. The authors have touched on a key principle of medical predictive analytics that is very interesting and relevant. While binary outcomes are oftentimes of utmost importance, it is nearly impossible to perfectly classify complex, real-world data in a binary fashion. As such, we agree that calibration is a very relevant aspect of all predictive analytical questions that is rarely touched upon and deserves more traction.

While we did not initially record calibration data from our previously published data set, we have since gone back and re-run our model with this particular question in mind. After re-running our model, we found our Brier score to be 0.045. Similarly, our calibration curve is shown in Fig. 1. As is typical for such curves, perfect calibration suggests a perfect alignment to the diagonal, whereas conservative predictions are shown in the upper quadrant and more aggressive predictions in the lower.¹ Our model was trained initially using a final layer built upon a probabilistic model (sigmoid). These types of models generally tend to have fewer issues with calibration as a result when compared

FIG. 1. Calibration curve for our model. Perfect calibration is suggested by following the *diagonal dashed line*. Observations above the diagonal suggest more conservative predictions, whereas those below the diagonal suggest more aggressive predictions. Figure is available in color online only.

with various other forms of classification models (such as support vector machines or other nonlinear methods that may be nonprobabilistic).²

It is, however, with great interest that, while we do note a discrepancy and separation between patients predicted to need readmission and those predicted not to require readmission, the model fails to calibrate most for true positives. As such, our model tends to predict readmissions conservatively (as is reflected in our calibration curve [Fig. 1]) with average likelihood probabilities for true readmissions reaching roughly 36%. While this still allows for a large discrepancy and separation of classes from those predicted to not require readmission (with average likelihood percentages reaching only approximately a 7% likelihood), it does shed light on the fact that the variables used in our model likely only play a relatively small role in the overall recipe for readmission. This reflects the realistic nature of hospital readmissions—there are many outside and confounding factors involved that likely contribute outside of observable clinical attributes. Despite our model's best discernment, it continues to predict readmissions on a case-by-case basis for even the strongest candidates at a modest probability of only 36% on average.

These probabilities, while relatively small, do in fact still provide useful clinical utility in classifying patients as potentially high risk. Discrimination continues to allow for quick and real-time risk stratification. Similarly, while 36% is a relatively small probability for patients testing positive, readmissions themselves are very costly and problematic. Similar to the argument and use of preventative medicine, spending a minimal amount of resources upfront despite a small underlying probability oftentimes tends to prove more cost-effective. Likewise, the large financial and emotional burden for patients required to return to the hospital is difficult for all parties involved and, if possible, should be prevented and avoided.

In conclusion, we agree with the need for more publicity of calibration as it relates to predictive analytics and as it continues to be a relevant aspect in clinical management. However, while useful, calibration needs be used

cautiously and in the appropriate context as failure to treat or overtreatment can often be disproportionately costly in case an unlikely event does occur.

Benjamin S. Hopkins, BS
Nader S. Dahdaleh, MD
Northwestern University, Chicago, IL

References

1. Brownlee J. How and when to use a calibrated classification model with scikit-learn. <https://machinelearningmastery.com/calibrated-classification-model-in-scikit-learn/>. Accessed January 30, 2020.

2. Kull M, Silva Filho TM, Flach P. Beyond sigmoids: how to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electron J Stat*. 2017;11(2):5052–5080.

INCLUDE WHEN CITING

Published online February 21, 2020; DOI: 10.3171/2020.1.SPINE191511.

©AANS 2020, except where prohibited by US copyright law