Supplemental Online Content

Marwaha JS, Chen HW, Habashy K, Choi J, Spain DA, Brat GA. Appraising the quality of development and reporting in surgical prediction models. *JAMA Surg.* Published online November 30, 2022. doi:10.1001/jamasurg.2022.4488

eMethods. Cohort Selection, Data Collection, Methodologic Limitations, and Data Availability

eReferences

This supplemental material has been provided by the authors to give readers additional information about their work.

Cohort selection

We selected four general surgery journals with the highest SJR2 rankings in the subject area "Surgery" from 2018-2020.[1] We selected articles by screening the titles and abstracts of every original article published in every issue of these four journals from January 1, 2018 - August 31, 2021, and included only studies that described a surgical prediction model. Three authors (JSM, HC, KH) independently screened studies for potential inclusion and reconciled any differences in final cohort composition by consensus.

Data collection

The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) Statement is the guideline used by the Enhancing the Quality and Transparency of Health Research (EQUATOR) Network for studies that describe the development, validation, or updating of clinical prediction models, and is endorsed by many prominent publishers and journals.[2] The TRIPOD categorizes studies as one of four types: development and internal validation; external validation only; incremental value (i.e., updating the parameters of a previously-developed model based on new data); and development and external validation. It also categorizes models as performing either a diagnostic or prognostic task. Study types and prediction tasks for each paper were independently classified by three authors (JSM, HC, KH), and any differences in classification were reconciled by consensus.

The TRIPOD Statement is a 37-item checklist of best practices for reporting of information in these studies. Within each item are several adherence criteria that all must be met in order for the study to receive credit for that item. Each paper was scored according to the TRIPOD Statement recommendations: an item received a "0" if at least one adherence criterion was not met, or received a "1" if all adherence criteria were met.[3] The overall compliance of the study with the TRIPOD statement was defined as the number of fulfilled items divided by the total number of items applicable to that study. These definitions and processes are consistent with many other studies that have scored papers against the TRIPOD Statement as well.[4–7]

Not all 37 items were applicable to all study types; for example, items specifically addressing model development steps were not applicable to studies performing external validation or model updating only. Items were excluded from analysis if they were not applicable to that particular study type according to the TRIPOD Statement or applicable to less than five studies in the full cohort.[3] For each study type, the following items were excluded from analysis:

Development and internal validation: 5c, 6b, 7b, 10c, 10e, 11, 12, 13c, 14b, 17, 19a
External validation: 5c, 6b, 7b, 10a, 10b, 10e, 11, 14a, 14b, 15a, 15b, 17
Incremental value: 5c, 6b, 7b, 11, 17
Development and validation: 5c, 6b, 7b, 10e, 11, 14b, 17

Of note, when examining reporting of model specifications, the TRIPOD statement specifically notes reporting regression coefficients. However we found that regression coefficients were not always the most appropriate piece of information to report for every type of model. We interpreted this statement to allow for more flexible parameter reporting based on the type of model that was developed. For example, when random forest, deep learning, or other complex models were developed, we accepted relevant feature importance measures such as mean decrease in Gini impurity[8] and Shapley values.[9]

Two authors (HWC, KH) independently scored articles against all adherence criteria, and any scoring differences were reconciled by a third author who independently reviewed the article (JSM). Cohen's Kappa statistic for inter-rater reliability of adherence to individual TRIPOD items prior to reconciliation of scoring differences was 0.83, with 8.3% of scored items requiring reconciliation. All authors who participated in the screening and scoring of articles have a background in conventional statistical and/or machine learning methods, development and validation of surgical prediction models, and/or assessing the quality of published surgical analyses.[10–14] Comparison testing between groups was conducted using Pearson's chi-square test with a significance threshold of 0.05.

Methodologic limitations

This study was not a formal systematic review. However, all surgical prediction model studies from the top four general surgery journals (as defined by SJR2 rankings) were included by manual title and abstract review. Studies published in other general surgery journals or surgical subspecialty journals were not included. Studies published in non-surgical journals that describe surgery-related or perioperative prediction models - including ones published in notable journals such as the *New England Journal of Medicine*[15] and the *Journal of the American Medical*

Association[16] - were also not included. The quality of surgical prediction model development and validation in studies published by these other journals is unknown and cannot be inferred from our findings.

Our study is also limited by the level of granularity of some TRIPOD statement elements. For example, item 16 is a composite measure that requires reporting of model discrimination, confidence interval, model calibration results, and additional performance measures to receive credit. This made it difficult for us to retrospectively discern which specific element of performance measure reporting needs improvement once all studies had been scored. Our ability to measure the quality of a study was also limited by what the authors chose to report in their published manuscript and supplemental materials. For example, the authors may have performed hyperparameter tuning during model development but not described this step in the manuscript. However, poor reporting alone has been recognized as a significant barrier to translating academic discoveries into clinical practice.[17] Finally, poor compliance with some elements of the TRIPOD statement such as a statement on funding (item 22) may have been a function of the journal rather than the study's authors; for example, not all journals include an explicit funding statement in the final published manuscript when the authors state during the submission process that no funding was received for the study.

Data availability

The full dataset for this study has been made publicly available and can be viewed and downloaded here: https://github.com/jayson-marwaha/TRIPOD *eReferences*

1. Guerrero-Bote VP, Moya-Anegón F. A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*. 2012;6(4):674-688. doi:10.1016/j.joi.2012.07.001

2. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350. doi:10.1136/bmj.g7594

3. Heus P, Damen JAA, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open*. 2019;9(4):e025611.

4. Li B, Feridooni T, Cuen-Ojeda C, et al. Machine learning in vascular surgery: a systematic review and critical appraisal. *npj Digital Medicine*. 2022;5(1):1-10.

5. Prediction models for living organ transplantation are poorly developed, reported and validated: a systematic review. *J Clin Epidemiol*. Published online February 4, 2022. doi:10.1016/j.jclinepi.2022.01.025

6. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol*. 2021;138:60-72.

7. Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol*. 2022;22(1):1-13.

8. Feature importances with a forest of trees. scikit-learn. Accessed July 4, 2022. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

9. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. Published online May 22, 2017. doi:10.48550/arXiv.1705.07874

10. MarwahaJayson S, Beaulieu-JonesBrendin R, KennedyChris J, et al. Design, Implementation, and Clinical Impact of a Machine Learning–Assisted Intervention Bundle to Improve Opioid Prescribing. *NEJM Catalyst Innovations in Care Delivery*. Published online March 16, 2022. doi:10.1056/CAT.21.0477

11. Goodman ED, Patel KK, Zhang Y, et al. A real-time spatiotemporal AI model analyzes skill in open surgical videos. Published online December 14, 2021. doi:10.48550/arXiv.2112.07219

12. Mechanical Thrombectomy in Patients Presenting with NIHSS Score <6: A Safety and Efficacy Analysis. *J Stroke Cerebrovasc Dis*. 2022;31(3):106282.

13. Choi J, Gupta A, Kaghazchi A, Htwe TS, Baiocchi M, Spain DA. Citation Inaccuracies in Influential Surgical Journals. *JAMA Surg*. 2021;156(8):791-792.

14. Choi J, Patil A, Vendrow E, et al. Practical Computer Vision Application to Compute Total Body

Surface Area Burn. *JAMA Surgery*. 2022;157(2):129. doi:10.1001/jamasurg.2021.5848

15. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration. *N Engl J Med*. Published online November 11, 2020. doi:10.1056/NEJMsa2001090

16. Wijnberge M, Geerts BF, Hol L, et al. Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. *JAMA*. 2020;323(11):1052-1060.

17. Sounderajah V, Ashrafian H, Karthikesalingam A, et al. Developing Specific Reporting Standards in Artificial Intelligence Centred Research. *Ann Surg*. 2022;275(3):e547.