JOURNAL OF ABDOMINAL WALL SURGERY
EUROPEAN HERNIA SOCIETY

TYPE Opinion
PUBLISHED 21 September 2026
DOI 10.3389/jaws.2026.17526

OPEN ACCESS

*CORRESPONDENCE
M. Verdaguer-Tremolosa,
mireia.verdaguer@uab.cat

RECEIVED 03 August 2026
REVISED 04 August 2026
ACCEPTED 07 September 2026
PUBLISHED 21 September 2026

CITATION
Verdaguer-Tremolosa M, Mojal S, Rodrigues-Gonçalves V, Martínez-López MP and López-Cano M (2026) Large registries, rare outcomes and neutral/inconclusive machine learning results in abdominal wall surgery. *J. Abdom. Wall Surg.* 5:17526. doi: 10.3389/jaws.2026.17526

Large registries, rare outcomes and neutral/inconclusive machine learning results in abdominal wall surgery

M. Verdaguer-Tremolosa[1,2]*, Sergi Mojal[3], V. Rodrigues-Gonçalves[1,2], M. P. Martínez-López[1,2] and M. López-Cano[1,2]

[1]Department of Surgery, UD of Medicine of Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain, [2]General and Digestive Surgery Department, Abdominal Wall Surgery Unit, Hospital Universitari Vall d'Hebron, Barcelona, Spain, [3]Institut d'Investigació Biomèdica Sant Pau (IIB-Sant Pau), Barcelona, Spain

KEYWORDS
abdominal wall surgery, Artificial intelligence, machine learning, registry, scientific publishing

Introduction

Artificial intelligence (AI) has become one of the most influential developments in contemporary surgical research, although many surgeons still have limited training in how these tools work and how their results should be interpreted. Machine learning (ML) is increasingly used to develop clinical prediction models, fuelled by the rapid expansion of prospective surgical registries and growing expectations that computational algorithms will improve individualized decision-making. Abdominal wall surgery (AWS) has followed this trend, with several studies exploring AI-based prediction of recurrence, postoperative complications and other clinically relevant outcomes [1–6].

At first glance, national registries containing hundreds of patients appear ideally suited for ML. However, this assumption deserves careful examination. Prediction models do not learn from patient numbers alone, they learn from informative outcome events. In AWS, the outcomes that matter most (mesh infection, recurrence, major surgical site occurrences or severe postoperative morbidity) may be uncommon in certain conditions such as in inguinal hernia repair. Consequently, even apparently large registries may contain relatively little information for predictive modelling.

Our experience analysing the management of inguinoscrotal hernia in the Spanish EVEREG registry illustrates this paradox. Although our cohort included almost 750 inguinoscrotal hernia repairs, postoperative complications occurred in only 52 patients. Multiple algorithms, including random forests, support vector machines, neural networks, k-nearest neighbours and elastic-net regression were explored. We were surprised that none demonstrated clinically meaningful superiority. So, we asked ourselves: are we sometimes asking more from ML than our data can provide? Our experience suggested that the main limitation was not the choice of algorithm, but the amount of information available in the dataset.

We therefore argue that such findings should not be interpreted as negative. They are neutral (or more appropriately, inconclusive) because they identify circumstances in which currently available registry data are insufficient to support robust prediction. This experience is the starting point for this Opinion Article.

Large registries are not necessarily informative prediction datasets

Prediction studies require more than large cohorts. Effective sample size depends on the number of events, candidate predictors and intended model complexity [7, 8]. A registry may therefore be excellent for epidemiology or quality assurance while remaining insufficiently informative for prediction modelling.

The EVEREG experience highlights this principle. Although the registry represented an excellent national cohort, only a small proportion of patients experienced postoperative complications. ML therefore had relatively few patients with complications from which to learn stable and reproducible patterns. This does not reduce the value of surgical registries, but it reminds us that a good registry is not necessarily a good dataset for prediction.

Machine learning cannot compensate for limited information

Modern algorithms can detect complex relationships, but they cannot generate biological information that is absent. When clinically relevant outcomes are rare, increasing computational complexity cannot replace genuinely observed events. Importantly, this limitation is not exclusive to ML: if key clinical variables are missing, or if too few outcome events are available, any predictive analysis will be limited (ML or conventional statistical methods). Consequently, different ML algorithms may show similar performance simply because the dataset does not contain enough information to improve prediction further.

Likewise, class imbalance should not be confused with lack of information. Oversampling, undersampling and SMOTE (Synthetic Minority Oversampling Technique) facilitate optimisation but cannot replace new independent patients with real outcome events [9, 10]. Future advances in AWS prediction will therefore depend more on richer clinical variables, improved follow-up and collaborative registries than on increasingly sophisticated balancing techniques.

Validation, interpretation and clinical usefulness

Scarcity of outcome events also influences validation strategy. Conventional train-test splits may waste valuable information, whereas bootstrap procedures and repeated cross-validation often provide more efficient internal validation when events are limited [11, 12].

Model evaluation should not be judged only by accuracy or AUROC (receiver operating characteristic curve). Calibration, sensitivity, positive predictive value and precision–recall curves frequently provide more meaningful assessments of clinical usefulness in imbalanced datasets [13, 14]. Ultimately, prediction models should be judged by their capacity to improve decision-making rather than by statistical novelty alone.

Prediction science beyond machine learning

Current evidence indicates that ML does not consistently outperform logistic regression across clinical prediction studies [15–17]. Rather than competing approaches, regression and ML should be viewed as complementary tools. Logistic regression remains valuable because it is simple and easy to interpret, whereas ML models are worthwhile when their additional complexity leads to a clearer improvement in prediction.

Importantly, predictive modelling should not be confused with causal inference. Variables identified as influential by ML are not necessarily causal determinants of postoperative complications. Clinical interpretation therefore remains essential.

Neutral/inconclusive results are scientifically valuable

Perhaps the principal message of this article is that inconclusive ML analyses deserve greater recognition. Publication bias favours successful algorithms, whereas studies reporting modest performance are often overlooked. Yet these investigations frequently provide equally valuable methodological information because they define the current limits of prediction.

In our opinion, the principal conclusion of the EVEREG experience is not that ML failed. Rather, the registry contained insufficient information to support reliable prediction of uncommon complications. This should not discourage the use of ML, but it should help us design better studies and know when prediction is realistically possible.

Discussion

The next-generation of AWS registries should focus not only on increasing patient numbers, but also on collecting more complete and clinically relevant information, using consistent definitions, improving follow-up and increasing collaboration between centres.

These developments are likely to contribute more to future prediction performance than replacing one algorithm with another. Frameworks such as TRIPOD + AI and PROBAST + AI provide an excellent methodological foundation for this evolution [18, 19].

AI undoubtedly has an important future in AWS. Nevertheless, prediction performance will always be constrained primarily by the quality and biological richness of the available data. Large registries with rare outcomes remind us that prediction models learn from informative events rather than from patient numbers alone. Neutral machine learning results should therefore not be regarded as failed studies. A model with limited predictive performance should not necessarily be considered a failed study, it may simply show that the available data are not yet sufficient for reliable prediction and help us design better studies in the future.

ML does not simply reveal what algorithms can learn; it also reveals what our registries are currently unable to teach, and it helps us understand the limitations of our own data.

Author contributions

Conceptualization: ML-C, MV-T, MM-L, VR-G, and SM. Writing original draft: ML-C and MV. Writing review and editing: all authors. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

ML-C has received honoraria for consultancy work, lectures, travel support, and participation in review activities from BD, Medtronic, and Gore. He is also an unpaid member of the EHS Board and Editor-in-Chief of JAWS.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. A generative AI tool (Chat GPT-5.5) was used to assist with language editing, grammar, and clarity. The authors reviewed and approved all changes and remain fully responsible for the final content. No AI tool is listed as an author.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1. Morris MX, Rustom F, Chun B, Limon D, Raghavan N, Guo L, et al. Artificial intelligence in surgical research: transformative impacts and evolving ethical challenges. *Am Surg* (2026) 92:698–709. doi: 10.1177/00031348251409740

2. Loftus TJ, Altieri MS, Balch JA, Abbott KL, Choi J, Marwaha JS, et al. Artificial intelligence-enabled decision support in surgery: state-of-the-art and future directions. *Ann Surg* (2023) 278:51–8. doi: 10.1097/SLA.0000000000005853

3. Stam WT, Goedknegt LK, Ingwersen EW, Schoonmade LJ, Bruns ERJ, Daams F. The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: a systematic review. *Surgery* (2022) 171:1014–21. doi: 10.1016/j.surg.2021.10.002

4. Lima DL, Kasakewitch J, Nguyen DQ, Nogueira R, Cavazzola LT, Heniford BT, et al. Machine learning, deep learning and hernia surgery. Are we pushing the limits of abdominal core health? A qualitative systematic review. *Hernia* (2024) 28:1405–12. doi: 10.1007/s10029-024-03069-x

5. Vogel R, Mück B. Artificial intelligence-what to expect from machine learning and deep learning in hernia surgery. *J Abdom Wall Surg* (2024) 3:13059. doi: 10.3389/jaws.2024.13059

6. Yu W, Ma Y, Wu J, Zhang M, Yang C. Interpretable machine learning model predicts 1-year inguinal hernia risk after robot-assisted radical prostatectomy. *J Robot Surg* (2025) 19:564. doi: 10.1007/s11701-025-02723-5

7. Riley RD, Snell KI, Ensor J, Burke DL, Harrell JFE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* (2019) 38:1276–96. doi: 10.1002/sim.7992

8. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* (2020) 368:m441. doi: 10.1136/bmj.m441

9. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* (2013) 14:106. doi: 10.1186/1471-2105-14-106

10. Piccininni M, Wechsung M, Van Calster B, Rohmann JL, Konigorski S, van Smeden M. 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. doi: 10.1016/j.jbi.2024.104666

11. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* (2001) 54:774–81. doi: 10.1016/s0895-4356(01)00341-9

12. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* (2016) 69:245–7. doi: 10.1016/j.jclinepi.2015.04.005

13. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the achilles heel of predictive analytics. *BMC Med* (2019) 17:230. doi: 10.1186/s12916-019-1466-7

14. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* (2015) 10:e0118432. doi: 10.1371/journal.pone.0118432

15. Wang J, Tozzi F, Ashraf Ganjouei A, Romero-Hernandez F, Feng J, Calthorpe L, et al. Machine learning improves prediction of postoperative outcomes after gastrointestinal surgery: a systematic review and meta-analysis. *J Gastrointest Surg* (2024) 28:956–65. doi: 10.1016/j.gassur.2024.03.006

16. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* (2019) 110:12–22. doi: 10.1016/j.jclinepi.2019.02.004

17. Murali M, Mann J, Parbhoo S. Surgical precision? Cutting through the hype of AI-augmented peri-operative risk prediction. *Anaesthesia* (2025) 80:1177–81. doi: 10.1111/anae.16662

18. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* (2024) 385:e078378. doi: 10.1136/bmj-2023-078378

19. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* (2025) 388:e082505. doi: 10.1136/bmj-2024-082505