ANNALS OF SURGERY OPEN
OPEN

Comment on: "Integrating Human Intuition into Prediction Algorithms for Improved Surgical Risk Stratification"

Elisabeth E. Smith, BMedSci, BMBS,* Brenig L. Gwilym, MBBCh, PhD,†‡ Philip Pallmann, PhD,‡§ and David C. Bosanquet, MD, FRCS*‡

We read with interest the above article by Berrigan et al.[1] The authors undertook a systematic literature review, assessing existing work in which surgeon intuition and risk prediction models were combined to estimate surgical risk. They identified 2 studies, and their own study, which reported the predictive abilities of clinicians, risk prediction models, and combined 'clinician-risk prediction models', which comprise a risk prediction model with a measure of clinician estimate of risk. In 1 article, the combined model marginally outperformed both of its separate components (area under curve 0.77 vs 0.76).[2] In the other 2 studies, the risk prediction model worked best, and the addition of clinician intuition scores did not improve prediction ability.[3,4] The studies all included general ± abdominal surgery.

We would like to draw attention to another relevant article, not described by Berrigan et al.[1] The article reports the development and validation of the second version of the Surgical Outcome Risk Tool (SORT-2), which incorporates surgeon intuition into the original version of SORT,[5] and is used to predict 30-day mortality.[6] The study was conducted at over 200 centers in the United Kingdom, Australia, and New Zealand and included several procedures from multiple surgical specialties. Over 22,000 patients were included in the analyses. The clinician prediction of mortality was obtained by asking relevant anesthetists, intensivists, and/or surgeons presurgery: "What is the estimate of the perioperative team of the risk of death within 30 days?". They were then given 6 categorical responses to choose from (<1%, 1%–2.5%, 2.6%–5%, 5.1%–10%, 10.1%–50%, and >50%). The authors found that combining this subjective assessment with SORT (SORT-2), led to a significant improvement in the area under receiver operating characteristic curve versus subjective assessment alone and SORT alone ($P < 0.001$ and $P = 0.021$, respectively).

We examined SORT-1 and SORT-2 models in our own study, focusing on predicting outcomes after amputation surgery[7,8] as part of PrEdiction of Risk and Communication of outcomE following major lower limb amputation: a collaboratIVE study (PERCEIVE). PERCEIVE was an international multicentre collaborative observational study that evaluated healthcare professionals' and statistical models' performance in predicting several outcomes at both 30 days (including 30-day mortality) and 1 year after amputations. Predictions of outcomes were collected by members of the treating teams before surgery and compared with actual outcomes. A total of 41 centers collected predictions of 553 patients undergoing amputation, and a total of 15 risk prediction tools were evaluated. Only 1 tool, SORT-2, combined both clinician estimates of risk with a 'standard' risk prediction model.

When looking at the prediction of 30-day mortality rates, the 13 relevant tools'[9] performance ranged from poor to acceptable (C-statistic 0.548–0.789). SORT-2 was the best-performing tool, and the only one to outperform clinicians in terms of discrimination, calibration, and overall performance. Specifically, SORT-2 (C-statistic 0.774) performed better than both SORT-1 (C-statistic 0.716) and healthcare professionals (C-statistic 0.758) in predicting 30-day mortality. A finding in both our 30-day and 1-year data was that clinicians, overall, consistently overestimate patients' risks. The improved performance of SORT-2 compared with subjective intuition was attributable to the consistent downgrading of the subjectively perceived risk, akin to Wong et al.[6]

The value of adding clinician intuition to risk prediction tools varies between the studies. This difference may be because clinician accuracy varies depending on the outcome, or that small studies fail to demonstrate the value of clinician intuition due to a type-2 error. Our 2020 narrative synthesis of surgeons' perception of postoperative outcomes and risks found clinician intuition was often outperformed by risk prediction calculators.[10] This review found, like PERCEIVE, that surgeons consistently overestimated the risk of mortality, possibly due to the very human desire to have one's expectations exceeded. Various types of cognitive biases, including recall bias, confirmation bias, anchoring bias, overconfidence bias, and self-serving bias, are also likely to impact mortality and morbidity estimations. It is these factors that solidify a need for robust validated prediction tools.

Nevertheless, studies into other areas of medical practice have found that wholly unquantifiable and very subjective variables such as "gut feeling" ('gestalt') and "sense of reassurance", when added to a clinical prediction model that included signs and symptoms on presentation, significantly improved recognition of serious bacterial infection in febrile children presenting to the emergency department.[11] Furthermore, such tools that integrate nurse ± parent concern have already been adopted in pediatric units across the UK.[12]

*From the *Gwent Vascular Institute, Royal Gwent Hospital, Newport, UK; †School of Medicine, Cardiff University, Cardiff, UK; ‡South East Wales Vascular Network, University Hospital of Wales, Cardiff, UK; and §Centre for Trials Research, Cardiff University, Cardiff, UK.*

D.C.B. was chief investigator and B.L.G. and P.P. were co-investigators on the Health and Care Research Wales funded PERCEIVE study (RfPPB-19-1642). Other author declares that there is nothing to disclose.

The Centre for Trials Research, Cardiff University receives infrastructure funding from Health and Care Research Wales.

Reprints: Philip Pallmann, PhD, Centre for Trials Research, Cardiff University, Neuadd Meirionnydd, Heath Park Way, Cardiff, CF14 4YS, Wales, UK. E-mail: pallmannp@cardiff.ac.uk.

Annals of Surgery Open (2024) 3:e485

Received: 10 July 2024; Accepted 12 July 2024

Published online 28 August 2024

DOI: 10.1097/AS9.0000000000000485

Assuming the value of clinician intuition, the question arises as to how the best clinician intuition can be captured. The 3 studies identified by Berrigan et al[1] measured clinician estimate of risk on a 100-point visual analog scale, on a 1–6 score of severity, and with a 3-point Likert scale ("lower than average risk", "average risk", or "higher than average risk"). In the pediatric articles mentioned above, merely the presence or absence of concern was captured. Which method is preferable is not known. Categorizing continuous data loses precision, which could affect how well intuition appears to impact risk prediction tools.

The timing of risk prediction should also be considered. In 1 study, surgeons were permitted to alter the scores postoperatively, which led to a small improvement in the model's performance when compared with the model in which preoperative scores were used.[2] Postoperative estimates of outcome are of some clinical value, but typically the most helpful aspect of risk prediction is in the preoperative process of shared decision-making.

Fundamentally, the aim of improving risk prediction methods for surgical procedures is to improve decision-making. Crucially, as it covers the entire breadth of medicine and surgery, important outcomes differ between different clinical problems and proposed treatments and differ between different patients with the same ailment. More research is needed to determine which outcomes need exploring within different specialities. We agree with a need for understanding how clinical judgment should be combined with models better; however, this should be a part of a larger effort to also understand how we use higher-quality information to make better-quality decisions.

REF0ERENCES

1. Berrigan MT, Beaulieu-Jones BR, Marwaha J, et al. Integrating human intuition into prediction algorithms for improved surgical risk stratification. *Ann Surg*. 2024;279:15–16.
2. Woodfield JC, Pettigrew RA, Plank LD, et al. Accuracy of the surgeons' clinical prediction of perioperative complications using a visual analog scale. *World J Surg*. 2007;31:1912–1920.
3. Farges O, Vibert E, Cosse C, et al. "Surgeons' Intuition" Versus "Prognostic Models": predicting the risk of liver resections. *Ann Surg*. 2014;260:923–928; discussion 928–930.
4. Marwaha JS, Beaulieu-Jones BR, Berrigan M, et al. Quantifying the prognostic value of preoperative surgeon intuition: comparing surgeon intuition and clinical risk prediction as derived from the American College of Surgeons NSQIP risk calculator. *J Am Coll Surg*. 2023;236:1093–1103.
5. Protopapa KL, Simpson JC, Smith NCE, et al. Development and validation of the Surgical Outcome Risk Tool (SORT). *Br J Surg*. 2014;101:1774–1783.
6. Wong DJN, Harris S, Sahni A, et al. Developing and validating subjective and objective risk-assessment measures for predicting mortality after major surgery: an international prospective cohort study. *PLoS Med*. 2020;17:e1003253.
7. Gwilym BL, Pallmann P, Waldron CA, et al. Short-term risk prediction after major lower limb amputation: PERCEIVE study. *Br J Surg*. 2022;109:1300–1311.
8. Gwilym BL, Pallmann P, Waldron CA, et al. Long-term risk prediction after major lower limb amputation: 1-year results of the PERCEIVE study. *BJS Open*. 2024;8:zrad135.
9. Preece R, Dilaver N, Waldron CA, et al. A systematic review and narrative synthesis of risk prediction tools used to estimate mortality, morbidity, and other outcomes following major lower limb amputation. *Eur J Vasc Endovasc Surg*. 2021;62:127–135.
10. Dilaver NM, Gwilym BL, Preece R, et al. Systematic review and narrative synthesis of surgeons' perception of postoperative outcomes and risk. *BJS Open*. 2020;4:16–26.
11. Urbane UN, Petrosina E, Zavadska D, et al. Integrating clinical signs at presentation and clinician's non-analytical reasoning in prediction models for serious bacterial infection in febrile children presenting to emergency department. *Front Pediatr*. 2022;10:786795.
12. Romaine ST, Sefton G, Lim E, et al. Performance of seven different paediatric early warning scores to predict critical care admission in febrile children presenting to the emergency department: a retrospective cohort study. *BMJ Open*. 2021;11:e044091.