HHS Public Access

Author manuscript

Surgery. Author manuscript; available in PMC 2022 April 01.

Published in final edited form as:

Surgery. 2021 April ; 169(4): 746–748. doi:10.1016/j.surg.2021.01.008.

Bridging the Artificial Intelligence Valley of Death in Surgical Decision-making

Jeremy Balch, MD[a], **Gilbert R. Upchurch Jr., MD**[a], **Azra Bihorac, MD**[b], **Tyler J. Loftus, MD**[a]

[a]Department of Surgery, University of Florida Health, Gainesville, Florida

[b]Department of Medicine, University of Florida Health, Gainesville, Florida

Artificial intelligence has underdelivered in improving health care. In 1970, a *New England Journal of Medicine* Special Article projected that computer science would augment or replace many intellectual functions of physicians.(1) Over the past 50 years, surgical literature has been similarly optimistic despite a paucity of high-level supporting evidence. The corresponding author may be guilty of perpetuating this disproportionate optimism. This article seeks to balance the discussion surrounding artificial intelligence in surgical decision-making by describing major barriers and potential solutions toward safe, effective clinical adoption of artificial intelligence-enabled decision-support platforms in surgery.

Artificial intelligence decision-support in surgery has largely failed to emerge from the valley of death: the chasm between model development and effective, real-world implementation. Medical knowledge is projected to double every 73 days; artificial intelligence is lauded for its ability to produce and operationalize medical knowledge by parsing large datasets and generating clinically useful predictions and classifications. Investigators are racing to produce artificial intelligence-enabled clinical decision-support tools, but few are implemented clinically, and even fewer change health care delivery or human behavior.(2, 3) Experienced surgeons are appropriately wary of hyped, novel solutions that are unsubstantiated by academic rigor and evidence of performance advantages in clinical settings. Even when intelligent algorithms accurately predict salient clinical outcomes hours or days in advance, these predictions will remain ineffectual unless they 1) establish trust in their accuracy, 2) target risk-sensitive decisions, and 3) integrate with clinical workflows. The purpose of this article is to describe mechanisms for bridging the artificial intelligence valley of death toward improved surgical care, as illustrated in Figure 1.

Building Accurate, Trustworthy Models

Surgical literature regarding artificial intelligence decision-support focuses almost exclusively on predictive accuracy, typically using single-institution data for model training

Please address correspondence to: Tyler J. Loftus, MD, Assistant Professor, Acute Care Surgery, University of Florida Health, Department of Surgery, Office phone: 352-273-5670, Cell phone: 864-888-7404, Fax: 352-273-5683, PO Box 100108, Gainesville, FL 32610-0108, tyler.loftus@surgery.ufl.edu.

Conflict of Interest Statement

The authors report no conflicts of interest.

and validation. Regrettably, direct comparisons between models are often hindered by heterogeneity in study design, patient populations, prediction windows, and performance metrics; adherence to reporting guidelines can address these challenges (e.g., SPIRIT-AI [Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence], CONSORT-AI [Consolidated Standards of Reporting Trials-Artificial Intelligence]). Most published literature does not compare model predictions with clinician predictions, but when they do, clinicians typically exhibit variable performance, while artificial intelligence models provide greater consistency and accuracy in risk assessments. (4, 5) By complying with reporting guidelines and making real-world comparisons between model and clinician predictions, it may be possible to improve the predictive performance of artificial intelligence decision-support in surgery. It is equally important to ensure that model outputs are trustworthy.

"Black box" algorithms must earn the trust of patients, clinicians, and investigators. This can be accomplished with model interpretation mechanisms that convey the relative importance or weight of input variables in determining outputs, thus indicating how and why predictions were made. Trust is also built by validating models externally and prospectively. External validation is facilitated by common data models that map similar variables from different institutions onto a single, interoperable scheme. Prominent examples include the open-source OMOP (Observational Medical Outcomes Partnership) common data model and the Fast Healthcare Interoperability Resource. Despite best intentions to maintain security in sharing data across institutions, all methods of data sharing risk privacy leakage and unintended discovery of protected health information. Alternatively, collaborative modeling without data sharing can be accomplished via federated learning, in which local models train separately and send gradients or coefficients to a global model. This approach optimizes data security and ensures generalizability across participating institutions.

Targeting Risk-sensitive Decisions

Model predictions are more likely to change patient and provider behavior when the risks and benefits of treatment options are complex, difficult to estimate, and evenly matched. For example, there is relative clinical equipoise in the decision for antibiotics alone versus operative source control of sepsis for an elderly patient with multiple comorbidities and non-perforated appendicitis with extensive peri-cecal inflammation. The morbidity of a laparoscopic appendectomy is relatively low but appendectomy may not be feasible in the setting of extensive peri-cecal inflammation, the morbidity of an open ileocecectomy is substantially higher, and the morbidity of failed non-operative management resulting in appendiceal perforation and abscess formation is also high. In this scenario, a model that accurately predicts outcomes for operative versus nonoperative management could meaningfully affect decision-making by providing the patient and surgeon with objectivity in the face of uncertainty.

Integrating with Clinical Workflows

Clinical application of artificial intelligence-enabled decision-support in surgery should occur only when there is adequate evidence of its safety and efficacy. This could be

accomplished by subjecting the algorithms to technology readiness level assessments, similar to those adopted by the American National Aeronautics and Space Administration (NASA) after the space shuttle Challenger tragedy.(6) It is also necessary to estimate deployment costs, or the organizational effort and resources required for clinical implementation. Deployment costs, though difficult to estimate and rarely reported, tend to be low when a model improves efficiency for a single, digital workflow, and high when the model affects multiple, non-digital workflows without a clear improvement in efficiency.(7)

When the technology is ready for implementation and deployment costs are low, artificial intelligence decision-support can yield impressive results, as demonstrated in randomized trials. Wijnberge et al.(8, 9) used an artificial intelligence algorithm to predict intraoperative hypotension, prompting Anesthesiologists to act early, more often, and differently, resulting in fewer intraoperative hypotensive episodes and less time-weighted hypotension. Shimabukuro et al.(10) deployed a machine-learning based sepsis prediction tool in medical-surgical ICUs, leading to shorter length of stay and decreased in-hospital mortality. Outside of clinical trial settings, successful implementation of artificial intelligence-enabled decision support requires not only integration with standard, digital workflows but defining and expanding the role of these workflows in surgical care.(11) Ideally, these algorithms would receive real-time electronic health record data inputs, including intraoperative data, and generate alerts and orders in an automated fashion.(12, 13)

Summary and Future Directions

If artificial intelligence-enabled decision support is to cross the valley of death and improve the lives of surgical patients and their caregivers, models must change human behavior and yield improved outcomes that outweigh model deployment costs. These goals are achievable by building accurate, reproducible models that target risk-sensitive decisions and integrate with digital workflows. Surgeons have a long, proud history of innovation and adoption of new technologies that offer performance advantages. Artificial intelligence in surgical decision-making must evolve and improve to earn its place in surgical care.

Funding Statement

AB was supported by R01 GM110240 from the National Institute of General Medical Sciences (NIGMS) and Sepsis and Critical Illness Research Center Award P50 GM-111152 from the NIGMS. TJL was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number K23 GM140268. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Rerefences

1. Schwartz WB. Medicine and the computer. The promise and problems of change. N Engl J Med. 1970;283(23):1257–64. [PubMed: 4920342]
2. Peterson ED. Machine Learning, Predictive Analytics, and Clinical Practice: Can the Past Inform the Present? JAMA. 2019.
3. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial Intelligence and Surgical Decision-Making. JAMA Surg. 2019.

4. Brennan M, Puri S, Ozrazgat-Baslanti T, Feng Z, Ruppert M, Hashemighouchani H, et al. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. Surgery. 2019;165(5):1035–45. [PubMed: 30792011]

5. Sauro KM, Soo A, de Grood C, Yang MMH, Wierstra B, Benoit L, et al. Adverse Events After Transition From ICU to Hospital Ward: A Multicenter Cohort Study. Crit Care Med. 2020;48(7):946–53. [PubMed: 32317594]

6. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG, Right Data Right Now C. Machine learning in intensive care medicine: ready for take-off? Intensive Care Med. 2020.

7. Morse KE, Bagley SC, Shah NH. Estimate the hidden deployment cost of predictive models to improve patient care. Nat Med. 2020;26(1):18–9. [PubMed: 31932778]

8. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, 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.

9. van der Ven WH, Veelo DP, Wijnberge M, van der Ster BJP, Vlaar APJ, Geerts BF. One of the first validations of an artificial intelligence algorithm for clinical use: The impact on intraoperative hypotension prediction and clinical decision-making. Surgery. 2020.

10. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res. 2017;4(1):e000234.

11. El Hechi MW, Nour Eddine SA, Maurer LR, Kaafarani HMA. Leveraging interpretable machine learning algorithms to predict postoperative patient outcomes on mobile devices. Surgery. 2020.

12. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, et al. MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery. Annals of surgery. 2019;269(4):652–62. [PubMed: 29489489]

13. Datta S, Loftus TJ, Ruppert MM, Giordano C, Upchurch GR, Rashidi P, et al. Added Value of Intraoperative Data for Predicting Postoperative Complications: The MySurgeryRisk PostOp Extension. Journal of Surgical Research. 2020;254:350–63.

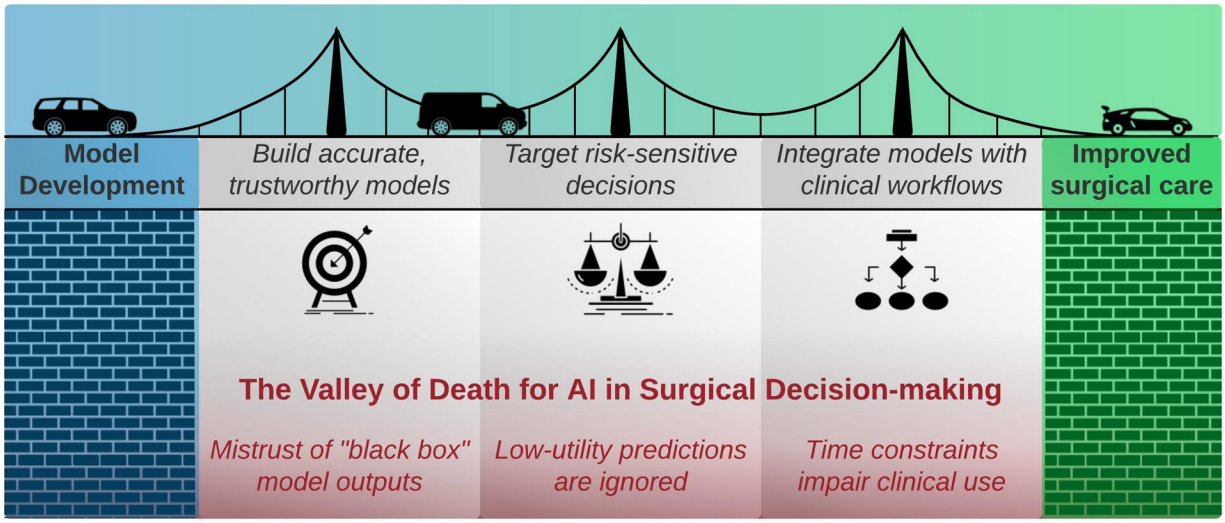

Figure 1:
To improve surgical care, artificial intelligence (AI)-enabled decision-support must overcome mistrust, low-utility predictions, and clinical time constraints.