



Published in final edited form as:

J Am Coll Surg. 2021 June ; 232(6): 971–972. doi:10.1016/j.jamcollsurg.2021.03.022.

Artificial Intelligence Assisted Surgical Quality Assessment: Hype or Hope?

Brian. T Bucher, MD, MS, FACS¹

¹Department of Surgery, University of Utah School of Medicine, Salt Lake City, UT

In this issue of the *Journal*, Zhu and colleagues present the external validation of a machine-learning (ML) based automated surveillance algorithm to detect surgical site infections (SSI) from the electronic healthcare record (EHR). (1) I commend the authors for undertaking an ambitious, technically challenging endeavor and bringing a rigorous approach to validating their approach. Overall, the authors conclude that SSI detection algorithms developed in one institution can generalize and be readily applicable to a second institution, thus giving a practical approach to accelerated chart reviews for surgical site infection detection.

To develop their algorithms, Zhu and colleagues built upon their previous work (2) and utilized electronic healthcare data from two geographically disparate academic medical centers, the University of Minnesota (UM) and the University of California at San Francisco (UCSF). The authors used the clinical data available in each institution's Enterprise Data Warehouse to abstract structured electronic health care data such as vital signs, laboratory and microbiology results, antibiotic administration, radiology procedures, and International Classification of Diseases Diagnosis Codes occurring between three and thirty days after the procedure. The authors then leveraged the high-quality outcome data available through each institution's American College of Surgeons' National Surgical Quality Improvement Program (NSQIP) as the reference standard for an SSI occurrence. (3) Using the data from one institution, the University of Minnesota, the authors developed three separate SSI detection algorithms for superficial, organ/space, or any-SSI outcome. After development, the algorithm's performance was validated in a blind fashion using a separate dataset from the UM (internal validation) and UCSF (external validation).

The authors demonstrate in external validation that the any-SSI algorithm had a sensitivity, specificity, and area under the receiver operating curve of 0.854, 0.734, 0.855, respectively. This classification performance was not significantly different from the performance during internal validation, thus demonstrating the algorithm's generalizability across institutions. Of note, the authors included patients whose follow-up was incomplete in the EHR, and the NSQIP surgical case reviewer had to contact to complete the 30-day follow-up. By including patients with incomplete EHR follow-up, the algorithm's performance if all patients had EHR data available is likely higher than reported. Given the low prevalence of SSI in the

CORRESPONDING AUTHOR Brian T. Bucher, MD, Assistant Professor of Surgery, Division of Pediatric Surgery, Department of Surgery, University of Utah School of Medicine, 100 North Mario Capecchi Drive, Suite #3800, Salt Lake City, UT 84113, Phone: 801-662-2950, Fax: 801-662-2980, brian.bucher@utah.edu.

CONFLICT OF INTEREST DISCLOSURE

The authors have no disclosures to report.

datasets, the algorithm's false-negative rate demonstrates the value of the author's approach to SSI surveillance. By accepting a false negative rate of 10% for missed SSI cases, using the author's algorithm, surgical quality assessment programs could reduce the burden of chart review to between 30 and 40% of eligible cases.

Despite the strong evidence of the ML models' generalizability, there are several limitations to the author's work. First, the authors utilized data from two large academic medical centers, and the generalizability of their algorithm to smaller, community hospitals is unknown. Second, the authors focused their validation on colorectal surgery procedures, which have a high prevalence of surgical site infections. The algorithm's classification performance to detect SSIs in lower prevalent procedures such as skin and soft tissue, orthopedic, neurosurgical procedures is unknown. Lastly, the authors utilized readily available structured data and did not take advantage of the rich data available in clinical notes using automated approaches such as natural language processing. (4, 5)

The skepticism aside, the authors are to be commended for addressing a technically challenging and complicated endeavor such as automated surgical site infection surveillance. The present work demonstrates the value of utilizing the rigorous clinical outcome data available through the American College of Surgeons' NSQIP program. This work will hopefully serve as a backbone for other AI-based surgical quality assessment surveillance systems.

Is AI-based surgical quality assessment surveillance ready for clinical use? Unfortunately, several technical challenges remain. To develop their algorithm, the authors performed significant feature engineering to create clinically relevant robust features. Mapping EHR data to these features is technically challenging to scale at present using existing EHR common data models. (6) Some of the algorithm's features are easily mapped through the EHR using standard terminologies and vocabularies, such as the LOINC terminology for laboratory values and vital signs. However, other features such as microbiology reports or imaging-based procedures do not have standard terminologies readily available, thus require custom onsite mapping at individual healthcare facilities. Second, the authors used antibiotic administration as an additional predictor. However, differences in medication formularies and facility-specific antibiotic prescription practices may impact these variables' performance during adoption. To address these barriers, best practices for ML implementation recommend validating the performance at each site before implementation. (7)

How can programs such as NSQIP leverage the author's findings to improve surgical quality assessment activities? The authors demonstrate that given the low prevalence of SSI, surveillance requires a significant amount of manual effort to perform the chart review process. The author's proposed approach can improve the manual chart review process's accuracy and efficiency by decreasing the number of operative events requiring manual review. Using an ML approach, the NSQIP program has the potential to expand to 100% sampling of all surgical procedures with confirmatory manual chart review for high-likelihood procedures identified through the surveillance algorithm. Smaller community hospitals, which do not have the resources available to perform intensive manual chart

review, could now participate in NSQIP surgical quality assessment activities assisted by an automated surveillance approach. Finally, as these surveillance algorithms become more accurate and generalizable, there is the potential to perform 100% autonomous surgical quality assessment through the EHR.

I congratulate the authors for proposing a novel SSI detection algorithm and validating this algorithm in an external data set. The author's approach highlights how the secondary use of the high-quality clinical data available through NSQIP can improve the efficiency and accuracy of surgical quality assessment activities. I look forward to the authors' future work in developing other ML-applications for surgical quality assessment.

ACKNOWLEDGEMENT

FUNDING/SUPPORT

Dr. Bucher is supported by grant 1K08HS025776 from the Agency for Healthcare Research

REFERENCES

1. Zhu Y, Simon GJ, Wick EC, et al. Applying Machine Learning Across Sites: External Validation of a Surgical Site Infection Detection Algorithm. *J Am Coll Surg*. 2021.
2. Skube SJ, Hu Z, Simon GJ, et al. Accelerating Surgical Site Infection Abstraction With a Semi-automated Machine-learning Approach. *Ann Surg*. 2020 Oct 14.
3. Ko CY, Hall BL, Hart AJ, Cohen ME, Hoyt DB. The American College of Surgeons National Surgical Quality Improvement Program: achieving better and safer surgery. *Jt Comm J Qual Patient Saf*. 2015 May;41(5):199–204. [PubMed: 25977246]
4. Bucher BT, Shi J, Ferraro JP, et al. Portable Automated Surveillance of Surgical Site Infections Using Natural Language Processing: Development and Validation. *Ann Surg*. 2020 Oct;272(4):629–36. [PubMed: 32773639]
5. Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*. 2011 Aug 24;306(8):848–55. [PubMed: 21862746]
6. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. *Journal of biomedical informatics*. 2016;64:333–41. [PubMed: 27989817]
7. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *New England Journal of Medicine*. 2019;380(14):1347–58.