Title

Applying Machine Learning Across Sites: External Validation of a Surgical Site Infection Detection Algorithm.

Journal

Journal of The American College of Surgeons, 232(6)

Authors

Abe-Jones, Yumiko
Najafi, Nader
Sheka, Adam
et al.

Publication Date

2021-06-01

DOI

10.1016/j.jamcollsurg.2021.03.026

Peer reviewed

HHS Public Access

Author manuscript

J Am Coll Surg. Author manuscript; available in PMC 2022 June 01.

Published in final edited form as:

J Am Coll Surg. 2021 June ; 232(6): 963–971.e1. doi:10.1016/j.jamcollsurg.2021.03.026.

Applying Machine Learning Across Sites: External Validation of a Surgical Site Infection Detection Algorithm

Ying Zhu, PhD, **Gyorgy J Simon, PhD**, **Elizabeth C Wick, MD, FACS**, **Yumiko Abe-Jones, MS**, **Nader Najafi, MD**, **Adam Sheka, MD**, **Roshan Tourani, PhD**, **Steven J Skube, MD**, **Zhen Hu, PhD**, **Genevieve B Melton, MD, FACS, PhD**

Institute for Health Informatics (Zhu, Simon, Tourani, Hu, Melton) and the Departments of Medicine (Simon) and Surgery (Wick), University of Minnesota, Twin Cities; Minneapolis, MN; and the Departments of Surgery (Abe-Jones, Najafi) and Medicine (Sheka, Skube, Melton), University of California San Francisco, San Francisco, CA

Abstract

BACKGROUND: Surgical complications have tremendous consequences and costs. Complication detection is important for quality improvement, but traditional manual chart review is burdensome. Automated mechanisms are needed to make this more efficient. To understand the generalizability of a machine learning algorithm between sites, automated surgical site infection (SSI) detection algorithms developed at one center were tested at another distinct center.

STUDY DESIGN: NSQIP patients had electronic health record (EHR) data extracted at one center (University of Minnesota Medical Center, Site A) over a 4-year period for model development and internal validation, and at a second center (University of California San Francisco, Site B) over a subsequent 2-year period for external validation. Models for automated NSQIP SSI detection of superficial, organ space, and total SSI within 30 days postoperatively were validated using area under the curve (AUC) scores and corresponding 95% confidence intervals.

RESULTS: For the 8,883 patients (Site A) and 1,473 patients (Site B), AUC scores were not statistically different for any outcome including superficial (external 0.804, internal [0.784, 0.874] AUC); organ/space (external 0.905, internal [0.867, 0.941] AUC); and total (external 0.855, internal [0.854, 0.908] AUC) SSI. False negative rates decreased with increasing case review volume and would be amenable to a strategy in which cases with low predicted probabilities of SSI could be excluded from chart review.

Correspondence address: Genevieve B Melton, MD, FACS, PhD, Colon and Rectal Surgery, 420 Delaware St SE, Mayo Mail Code 450, Minneapolis, MN 55455. gmelton@umn.edu.

Author Contributions

Study conception and design: Zhu, Simon, Wick, Hu, Melton

Acquisition of data: Abe-Jones, Najafi, Sheka, Skube

Analysis and interpretation of data: Zhu, Simon, Abe-Jones, Najafi, Tourani, Skube, Hu, Melton

Drafting of manuscript: Zhu, Simon, Wick, Abe-Jones, Najafi, Sheka, Tourani, Skube, Hu, Melton

Critical revision: Zhu, Simon, Wick, Abe-Jones, Najafi, Sheka, Tourani, Skube, Hu, Melton

Disclosure Information: Nothing to disclose.

CONCLUSIONS: Our findings demonstrated that SSI detection machine learning algorithms developed at 1 site were generalizable to another institution. SSI detection models are practically applicable to accelerate and focus chart review.

Surgical site infections (SSIs) account for 20% of all infections in the hospital setting. These infections are highly morbid, increase hospital length of stay, dramatically increase the risk of mortality, and on average, increase the cost of hospitalization by more than $20,000 per patient.[1] Because many SSIs are preventable, they are included as elements of Centers for Medicaid and Medicare pay-for-performance programs, including the Hospital-Acquired Condition Reduction and the Hospital Value-Based Purchasing Programs; resulting in low-performing hospitals being at risk of losing billions of dollars in annual revenue.[2,3] Given the significant consequences of SSI on individual patients and their aggregate impact on healthcare systems financially, careful tracking of SSI outcomes and other hospital-acquired infections (HAIs) is a national imperative in conjunction with continuous quality improvement around prevention.

Unfortunately, tracking SSIs and other surgical complications is costly and resource-intensive today. Although national surgical registries, including the American College of Surgeons NSQIP, Society for Thoracic Surgeons National Database, National Trauma Data Bank, and National Healthcare Safety Network, among others, provide high quality data on short-term surgical outcomes, most of these databases require significant resources in the form of trained personnel to manually abstract these outcomes using retrospective chart review. Data collection is, therefore, a time-consuming and expensive process. In most cases, larger hospitals only track data on a sample of patients rather than all because of this cost. Moreover, program and data collection costs for NSQIP and other outcomes registries bias participation toward larger academic medical centers.[4]

To improve the efficiency of surgical adverse event reporting, investigators have started to explore the feasibility of using automated solutions in place of manual chart review. In most cases, this process has relied heavily on administrative data and claims data.[5,6] In some cases, narrative clinical notes and other electronic health record (EHR) data have also been explored.[7–10] However, these studies have been carried out at a single institution, and performance of these solutions varies widely across the studies.

In our previous work, we developed postoperative SSI detection algorithms based on structured EHR data for overall, superficial, and organ-space SSI using NSQIP as a gold standard comparator.[11] As opposed to algorithms useful for prediction of SSI, this algorithm is used for detection of SSI, which has occurred using variables associated with diagnosis and confirmation (eg imaging studies, antibiotic therapy). Our machine learning models demonstrated high specificity and high negative predictive values, making them potentially useful in practice to reduce the workload and increase efficiency of clinical data abstractors by several fold. Also, when we applied these machine learning models to more contemporaneous surgical cases, we observed stable and robust performance over time at the same center.

An outstanding question remains as to whether machine-learning algorithms, such as the ones described developed at a single site, are generalizable to other sites. This study was designed to evaluate the performance of EHR-derived postoperative SSI detection models at an outside, geographically separate site. This external validation also serves as a test for determining the feasibility of applying these models into widespread practice. We hypothesized that representation of variables used in a model of EHR-based clinical indicators to detect SSI would be consistent across sites, making automated detection models transferrable from 1 institution to another.

METHODS

Setting

This study was performed on cohorts of surgical patients at 2 independent NSQIP-participating hospitals: University of Minnesota Medical Center (UMMC, Site A) located in Minneapolis, MN and University of California San Francisco Medical Center (UCSF, Site B) in San Francisco, CA. Institutional review board approval (IRB) was obtained from both institutions (multi-institutional IRB), with individual informed consent waived for this minimal risk study.

Dataset and outcomes

Figure 1 illustrates the overview of the methodological approach used in this study. Site A data were collected between 2011 and 2014. Site B data were collected between 2016 and 2017. At each site, structured EHR data for all patients in that institution's NSQIP registry during the study period were extracted from clinical data repositories prospectively maintained by each institution. The postoperative SSI outcome gold standard data were obtained from the NSQIP registry, using each patient's unique identifier and date of surgery to link the 2 datasets. Per NSQIP guidelines, patients with infection present at the time of surgery (PATOS) were included. SSI determinations were obtained using standard NSQIP data collection methods, including contacting the patient. Patient contact was not recorded in the EHR by the NSQIP abstractor. While the Site A cohort used a classical sampling methodology for patients included in NSQIP registry database and did expand further to neurosurgery, urology, and gynecology between 2011 and 2014, the Site B cohort included all colorectal surgery cases in an effort to perform additional quality improvement around colorectal SSI.

Structured EHR data—From each EHR data repository, we collected a range of data elements (eTable 1). This included demographic information (eg age, sex), relevant laboratory results (eg white blood cell count, hemoglobin, glucose, creatinine, troponin) including microbiological results (eg urine culture, blood culture, wound culture), relevant diagnosis codes (eg International Classification of Diseases [ICD] diagnosis codes for SSI, urinary tract infection, and sepsis with mappings between ICD9 and ICD10 through intelligent medical objects interface terminology), orders and procedures for making a diagnosis or providing treatment (eg CT-guided interventional radiology drainage to treat SSI, an order for a blood culture), vital signs (eg temperature, heart rate, blood pressure) and

administration of antibiotics. We also extracted the American Society of Anesthesiologists (ASA) physical status classification and surgical wound classification.

For this SSI detection algorithm, clinical data of interest were included during the window from postoperative days 3 to 30. This was done to account for a recovery period over the first 2 postoperative days, when some abnormal measurements are common. During the postoperative window, repeated values (ie laboratory result and vital measurement) are summarized by extreme (lowest and highest) and average values. Binary variables were created for relevant orders, procedures, and diagnosis codes for their presence or absence. For microbiology tests, binary variables were created in addition to identification of specific bacterial types (eg *Escherichia coli, Klebsiella pneumonia, Enterobacter, Enterococcus*) and bacterial morphologies (eg gram positive rods, gram negative cocci).

NSQIP SSI gold standard—As previously described,[12] NSQIP SSI data were collected by trained surgical clinical reviewers over the 30-day postoperative window. When this process has undergone a data quality auditing process, it has been reported to have over 95% accuracy and reproducibility.

Data analysis

Once EHR data and NSQIP data were linked, data preprocessing and modeling were tuned and optimized in order to better serve the purpose of model generalizability across sites. For example, preprocessing missing data imputations, we used the best practices proposed in Hu and colleagues[13] including imputation schemes customized to each variable. Missing data in the variables with repeated measurements (ie labs and vitals) were imputed using the average value of patients without any adverse events in order to minimize bias. For ASA and wound classification, missing values were imputed using the median of all cases, as missingness was assumed to be random. Each SSI outcome was modeled independently on the UMMC dataset using lasso logistic regression, with its penalty parameter determined through the 10-fold cross validation process. With respect to model optimization for transferability, before fitting the lasso-penalized logistic regression model, we applied additional steps to reduce the dimensionality of the analytical matrix by using causal variable screening (PC-Simple algorithm) followed by backward elimination, which could prevent an overfitting issue during external validation. Before modeling, the high dimensionality of independent variables were first reduced by causal variable screening[14] using PC-Simple algorithm[15] with a maximal condition set size of 3,[16] followed by backwards elimination with a significance level of 0.05. By doing so, only the features most relevant to the outcomes are selected and included in the final model, in order to minimize collinearity. No interaction terms were included in the model. All analyses were conducted using R version 3.5.1.

Evaluation

Model performance was evaluated by measuring the area under the ROC curve (AUC) score, sensitivity, and specificity. Youden’s index was used to determine the optimal cut-off to maximize the sum of sensitivity and specificity. First, the model was internally evaluated on a leave-out test set at Site A. To assess the variability of the detection performance,

1,000 bootstrap replications were performed on the Site A test set and the empirical 95% confidential interval (CI) of the evaluation metrics was reported. Next, the models were externally evaluated on the (entire) Site B data set: a single value for each metric was reported. In order to further assess cases incorrectly classified by the algorithm, manual review of cases with low predicted probability of SSI was performed by a single surgeon (SS), with attention on clinical notes such as progress notes and discharge summaries. If there was a question about whether there was an SSI or not, the case was reviewed with 2 other surgeons (AS, GM) to ensure consensus.

RESULTS

Table 1 summarizes patient cohorts from the 2 health systems including demographic information, case type, and SSI outcomes. The table also provides data and values for each of the algorithm model features that were significant in at least 1 of the models. Categorical variables are reported as the percentage of occurrence; continuous variables are reported as the median and the interquartile range. We observed that all variables that were features to the model were consistently represented with respect to most variable mappings. There was some additional variable mapping required around microbiology orders, antibiotic orders, and microbiology results to properly transfer the model. Overall, the Site B cohort has higher infection rates for all SSI types, likely due to the higher rates of colorectal and hepato-pancreato-biliary surgical cases. The cohorts also differed in age with Site B patients older, greater rates of inpatient cases, and higher wound class as compared to the Site A cohort.

Performance of each SSI detection model is summarized in Table 2. External validation of the model at Site B for superficial SSI demonstrated an AUC of 0.804, similar to Site A AUC internal validation (CI 0.784, 0.874). For organ/deep space SSI, external validation of AUC was 0.905, also with similar internal validation (CI 0.867 0.941). Finally, total SSI AUC external validation was 0.855, with internal validation AUC (CI 0.854, 0.908). Table 3 summarizes the significant variables from the models and the corresponding coefficients for the multivariate detection models developed with the Site A dataset for superficial, organ/space and the total SSI.

Figure 2 illustrates the tradeoff between false negative rate (FNR) and proportion of cases requiring manual review by illustrating the change of FNR against the case review volume needed for each model when applied internally on Site A dataset and externally on Site B dataset. On the horizontal axis it shows the FNR, and on the vertical axis it shows the percentage of cases requiring expert review. When the predicted probability of SSI falls below a predetermined threshold, the patient is automatically classified as a non-case (negative) and is excluded from further consideration. Patients above this threshold undergo manual review. Increasing automatic exclusion of patients reduces costly manual review, but increases the false negative rate (the proportion of SSI cases erroneously classified as negative). The model performs better generally on the Site A dataset, with fewer cases requiring review with relatively better performance as compared to Site B.

To further understand where the algorithm was less accurate, manual review was performed for Site A cases who were reported to have SSI by NSQIP, but had very low probabilities of SSI with our algorithm. This demonstrated that NSQIP data were nearly always correct. In most cases, these SSI cases were missed because they were mentioned in progress notes or discharge summaries, but discrete data signals were not entered or recorded as a discrete diagnosis. A large majority of these cases were also from earlier years of study; we have noted that this issue appeared to mostly resolve itself in more recent years, with EHR documentation and coding practice improvements, and that natural language processing methods to extract these signals from clinical notes into the detection algorithm may further improve model performance.[17,18]

DISCUSSION

As healthcare faces an affordability crisis becoming increasingly regulated and value-based, the need for automation and efficiencies in performing quality improvement are needed. Artificial intelligence algorithms and the application of machine learning holds significant promise for improving several of these processes. Unfortunately, algorithms often experience some level of degradation between populations and over time, as treatments and outcomes change, as well as due to differences in practice between sites and sets of surgeons. Best practices in transferring models between sites have not been well described, including adequate methods to deal with data preprocessing and model overfitting. In this study, we previously built and then validated machine learning models using structured EHR data for the detection of SSIs leveraging NSQIP outcomes. We then optimized these models for transferability and tested their performance externally at another site more contemporaneously over several years of data. Our results demonstrate that these models have good potential for reproducibility, and there is a wide range of similar opportunities for the application of these detection models in clinical practice, such as other HAIs and registries (eg Centers for Disease Control National Hospital Surveillance Network measures).

Broadly, an approach such as this may bring value to healthcare systems on several levels. While full automation of abstraction would be ideal, instead, the use of more focused abstraction may be more practical in keeping the accuracy of registry outcomes high while eliminating a large proportion of abstraction for cases that are clearly negative.[19] Additional benefits include increased capacity for abstracting additional cases, possible decreased abstraction costs, lowering the barrier for broader participation of institutions that do not currently participate, and potentially additional focus on quality improvement and feedback to surgeons. This would provide opportunities for deploying interventions targeting SSI rates and evaluating their effectiveness.

We also observed a similar tradeoff between FNR and the review burden in our analysis for both sites. In order to reach current NSQIP standards, which allow for interrater reliability disagreement rates to 5%, between 40% and 55% of cases would still need to be reviewed to achieve this degree of reliability, and when reliability standards are relaxed to 10% for organ/deep SSI, only 24% to 37% of cases require review. Overall, our findings suggest that with a large percentage of cases eliminated as negatives, the chart review process may be

considerably accelerated using the proposed approach. As illustrated in Figure 1, the optimal tradeoff for false positives and false negatives should be used to adapt the cutoff point according to the acceptability of each for the particular use case. This algorithm would be used primarily by surgical abstractors for NSQIP to automate case determination for those cases clearly positive and those clearly negative to decrease the number of cases requiring manual review by many-fold.[19]

Generally speaking, we observed good AUC scores for both internal and external validations with our algorithms. One interesting aspect of the external model validation to Site B were the differences in case make-up between sites. NSQIP has encouraged sites interested in quality improvement around certain aspects of care, such as prominently SSI, the opportunity to include all cases vs a sampling methodology. In this case, Site A was interested in involving additional disciplines in NSQIP, including gynecology oncology, neurosurgery, and urology for additional tracking of quality improvement. In contrast, Site B was interested in decreasing SSI rates and other major complications associated with colon and rectal surgery and hepatobiliary surgery. Therefore, we also observed that even though Site B has higher rates of positive SSI cases, likely due to the greater proportion of colon and rectal surgical cases as compared with Site A, where model development occurred, our SSI detection models still showed good detection performance across sites. The observed resilience of our algorithms despite variations in surgical populations and sites also demonstrates good robustness of this approach for the task of SSI detection. We did, however, also observe worse performance of our superficial SSI detection model. In fact, we observed an extended plateau for performance around superficial SSI indicating that signals for superficial SSI are more challenging. This is consistent with findings nationally, particularly with gaps in follow-up in the postoperative ambulatory setting, where care can be fragmented and patients can receive follow-up documented in other systems or with other providers.

There are several limitations of this study, including our validation at a single academic center as opposed to a variable set of centers with nonacademic/community surgical populations. Also, the study was conducted at 2 sites with the same third party EHR vendor, Epic Corporation. For this reason, the underlying data model and structure of clinical data may be more similar as compared to a health system with a different EHR platform. Despite the similarities in EHR vendor between sites, our models did require some additional feature engineering due to differences in local terminology used for imaging studies, microbiology results, and orders for inpatient care. These efforts will be accelerated by mapping of features to interoperable standards and data exchange frameworks (eg HL7 FHIR, certified health information technology data standard). Also, if applied to sites with large numbers of specialized patient populations (eg a center specializing in surgical oncology or orthopaedic procedures), performance of our model may not be generalizable. It is also possible that there were differences in routine practices for diagnosing or treating SSIs, which we did not take into account specifically in our model development. Finally, we expect that algorithms will drift over time and require ongoing maintenance as care practices and outcomes change. Our future work will expand upon this work by developing and validating similar detection models for other HAIs including pneumonia, urinary tract infection, sepsis, and septic shock.

CONCLUSIONS

Overall, our machine learning approach for SSI detection generalized well to an outside site using its EHR data over a 2-year period. The resulting AUC scores demonstrate that these approaches have very good potential for aiding in regulatory reporting and the process of data abstraction and ultimately, for aiding with work in clinical quality improvement. Scaling our approach to a wide range of sites and maturing our approach to multiple EHR platforms using clinical standards to interoperate is our next step. Ultimately, the maturation of artificial intelligence for clinical care has the potential to have a positive impact on surgical care and surgical quality improvement.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Support:

This research was supported by the University of Minnesota Academic Health Center Faculty Development Award, Agency for Healthcare Research and Quality (R01HS24532), NIH Clinical and Translational Science Award program (UL1TR000114), NIH National Institute of General Medical Sciences (R01GM120079), Fairview Health Services, University of Minnesota Physicians, and University of California, San Francisco Medical Center.

Abbreviations and Acronyms

AUC	area under the curve
EHR	electronic health record
FNR	false negative rate
HAI	hospital-acquired infections
SSI	surgical site infections
UMMC	University of Minnesota Medical Center

REFERENCES

1. Ban KA, Minei JP, Laronga C, et al. American College of Surgeons and Surgical Infection Society: Surgical Site Infection Guidelines, 2016 Update. J Am Coll Surg 2017;224:59–74. [PubMed: 27915053]
2. Centers for Medicare and Medicaid Services. Available at: https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf. Accessed April 8, 2021.
3. Centers for Medicare and Medicaid Services. Available at: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Downloads/HAC-Reduction-Program-Fact-Sheet.pdf. Accessed April 8, 2021.
4. Weiss A, Anderson JE, Chang DC. Comparing the National Surgical Quality Improvement Program with the Nationwide Inpatient Sample Database NSQIP vs NIS DatabaseLetters. JAMA Surg 2015;150:815–816. [PubMed: 26061977]

5. Mu Y, Edwards JR, Horan TC, et al. Improving risk-adjusted measures of surgical site infection for the National Healthcare Safely Network. Infect Control Hosp Epidemiol 2011;32:970–986. [PubMed: 21931247]
6. Levine PJ, Elman MR, Kullar R, et al. Use of electronic health record data to identify skin and soft tissue infections in primary care settings: a validation study. BMC Infect Dis 2013;13:171. [PubMed: 23574801]
7. Tinoco A, Evans RS, Staes CJ, et al. Comparison of computerized surveillance and manual chart review for adverse events. J Am Med Inform Assoc 2011;18:491–497. [PubMed: 21672911]
8. Wu ST, Sohn S, Ravikumar KE, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Ann Allergy Asthma Immunol 2013;111:364–369. [PubMed: 24125142]
9. Yetisgen M, Klassen P, Tarczy-Hornoch P. Automating data abstraction in a quality improvement platform for surgical and interventional procedures. Academy Health 2014; 2:1114.
10. Murff HJ, FitzHenry F, Matheny M, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306:848–855. [PubMed: 21862746]
11. Hu Z, Simon GJ, Arsoniadis EG, et al. Automated detection of postoperative surgical site infections using supervised methods with electronic health record data. Stud Health Technol Inform 2015;216:706–710. [PubMed: 26262143]
12. National Surgical Quality Improvement Program. Available at: https://www.facs.org/~/media/files/qualityprograms/nsqip/nsqip_puf_userguide_2017.ashx. Accessed April 8, 2021.
13. Hu Z, Melton GB, Simon GJ. Strategies for handling missing data in detecting postoperative surgical site infections. ICHI 2015;2015:499.
14. Ma S, Statnikov A. Methods for computational causal discovery in biomedicine. Behaviormetrika 2017;44:165–191.
15. Kalisch M, Machler M, Colombo D, et al. Causal inference using graphical models with the R package pcalg. Available at: https://www.jstatsoft.org/v047/i11. Accessed April 8, 2021.
16. Aliferis CF, Tsamardinos I, Statnikov A. HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA Annual Symp Proc 2003;2003:21–25.
17. Skube SJ, Hu Z, Arsoniadis EG, et al. Characterizing surgical site infection signals in clinical notes. Stud Health Technol Inform 2017;245:955–959. [PubMed: 29295241]
18. Liu H, Sohn S, Murphy S, et al. Facilitating post-surgical complication detection through sublanguage analysis. AMIA Jt Summits Transl Sci Proc 2014;77–82. [PubMed: 25717405]
19. Skube SJ, Hu Z, Simon GJ, et al. Accelerating surgical site infection abstraction with a semi-automated machine-learning approach. Ann Surg 2020 10 14 [Online ahead of print].

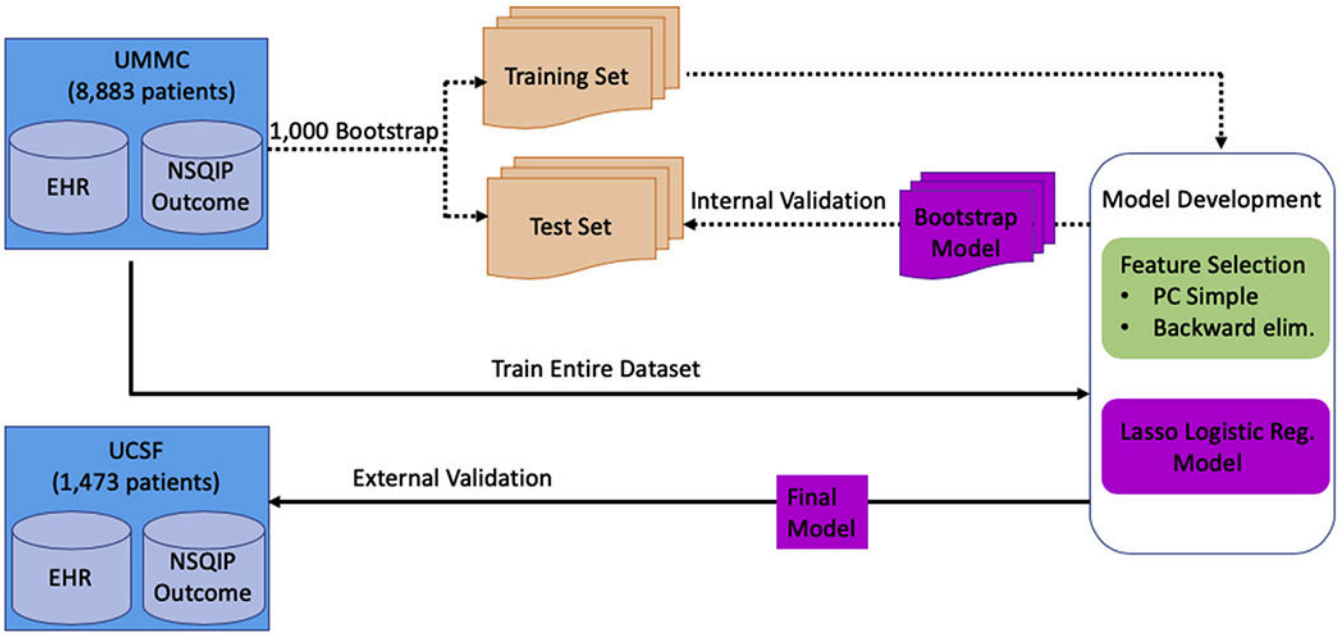

Figure 1.

Overview of experimental approach for internal and external validation of algorithms for surgical site infection detection. EHR, electronic health record; UCSF, University of California, San Francisco; UMMC, University of Minnesota Medical Center.

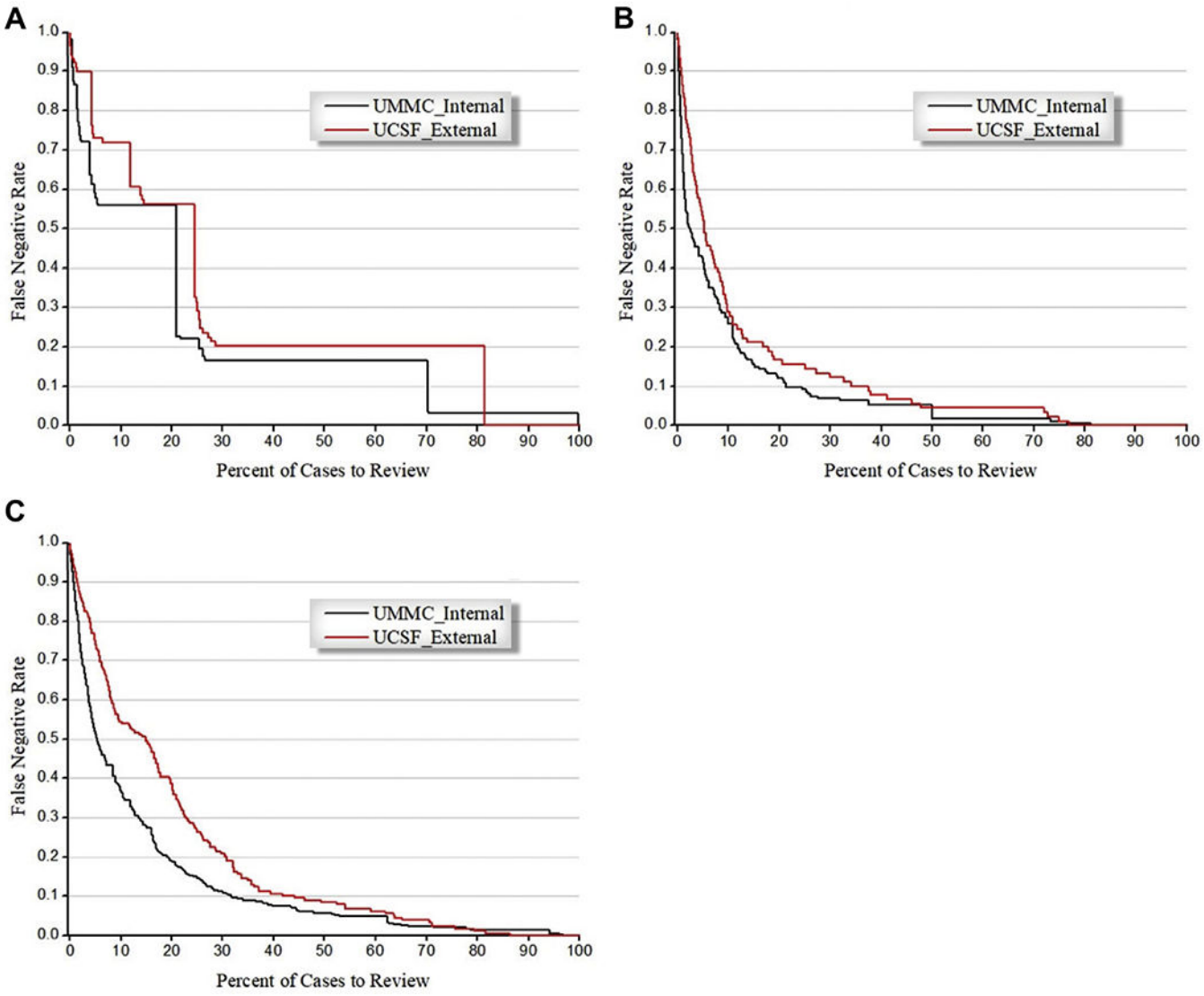

Figure 2.

False negative rate (FNR) of 3 models vs percentage of case reviews based on predicted probabilities of surgical site infection (SSI) development. (A) Superficial SSI; (B) organ space SSI; (C) total SSI. FNR is calculated by applying the models internally on University of Minnesota Medical Center (UMMC, Site A) dataset and externally on University of California, San Francisco (UCSF, Site B) dataset. Generally, the results suggest that the developed models have lower FNR on the Site A dataset.

Table 1.

University of Minnesota Medical Center (Site A) and University of California, San Francisco (Site B) Surgical Cohorts Summarized

Characteristic	Site A	Site B
Total surgical case, n	8,883	1,473
Demographic		
Age, y, median (range)	53 (39, 63)	61 (48, 69)
Sex, m, n (%)	3,666 (41.27)	698 (47.39)
Procedure type, n (%)		
Colon and rectal	613 (6.90)	379 (25.73)
Hepatopancreatobiliary	336 (3.78)	242 (16.43)
Other	7,934 (89.32)	852 (57.84)
SSI outcomes, n (%)		
Superficial	225 (2.53)	89 (6.04)
Organ/space	174 (1.96)	90 (6.11)
Total	487 (5.48)	178 (12.08)
Algorithm model feature		
Inpatient/outpatient (outpatient), n (%)	2,994 (33.70)	256 (17.38)
Surgical wound classification, at least level 2, clean-contaminated, n (%)	3,309 (39.54)	903 (61.30)
SSI-related imaging treatment study ordered, n (%)	753 (8.48)	94 (6.38)
SSI-related procedure performed, n (%)	2,062 (23.21)	75 (5.09)
SSI-related ICD diagnosis code recorded, n (%)	344 (3.87)	284 (19.28)
Fluid culture ordered (fluid), n (%)	214 (2.41)	32 (2.17)
Abscess culture ordered (abscess), n (%)	76 (0.86)	11 (0.75)
Wound culture ordered (wound), n (%)	168 (1.89)	6 (0.41)
Microbiology test positive (corynebacterium), n (%)	28 (0.32)	3 (0.20)
Microbiology test positive (enterococcus), n (%)	179 (2.02)	29 (1.97)
Microbiology test positive (staphylococcus), n (%)	358 (4.03)	16 (1.09)
Microbiology test positive (streptococcus), n (%)	107 (1.21)	3 (0.20)
Superficial SSI antibiotic ordered (oral), n (%)	2,235 (25.16)	189 (12.83)
Organ SSI antibiotic ordered (intravenous), n (%)	2,228 (25.08)	190 (12.90)

Characteristic	Site A	Site B
Minimum partial thromboplastin time, median (range)	32 (29, 37)	26.8 (25.2, 29.2)
Minimum systolic blood pressure, median (range)	111 (102,123)	109 (97, 122)
Minimum temperature, median (range)	97.9 (97.3, 99.1)	97.3 (96.8, 97.7)

SSI, surgical site infection.

Table 2.

Model Performance Metrics Calculated under Internal and External Validation

	Site A: internal validation			Site B: external validation		
SSI outcomes	**AUC (95% CI)**	**Sensitivity* (95% CI)**	**Specificity* (95% CI)**	**AUC**	**Sensitivity***	**Specificity***
Superficial SSI	0.784, 0.874	0.744, 0.900	0.715, 0.808	0.804	0.798	0.747
Organ/space SSI	0.867 0.941	0.736, 0.967	0.724, 0.918	0.905	0.789	0.905
Total SSI	0.854, 0.908	0.776, 0.913	0.744, 0.866	0.855	0.854	0.734

Site A: University of Minnesota Medical Center. Site B: University of California San Francisco.

*Calculated using cut-off point decided by Youden's index (maximizing the sum of sensitivity and specificity).

AUC, area under the curve; SSI, surgical site infection.

Table 3.

Significant Variables and Coefficients of Models for Superficial, Organ, and Total Surgical Site Infection Detection

Variable	Superficial	Organ/space SSI	Total SSI
(Intercept)	−4.602	−16.864	−3.917
Outpatient	−0.157	–	−0.119
Wound classification	–	0.037	0.040
SSI-related imaging	–	1.213	–
SSI-related procedure	–	–	0.374
SSI-related ICD	0.735	0.487	1.213
Fluid	–	1.098	0.768
Abscess	0.274	2.134	2.254
Wound	2.156	–	1.790
Corynebacterium	0.455	–	–
Enterococcus	–	0.961	–
Staphylococcus	–	–	0.558
Streptococcus	–	0.702	0.305
Superficial SSI antibiotic	1.256	–	1.442
Organ SSI antibiotic	–	1.106	–
Minimum partial thromboplastin time	0.007	–	–
Minimum systolic blood pressure	–	−0.009	−0.001
Minimum temperature	–	0.128	–

SSI, surgical site infection.