Prospective evaluation of data-driven models to predict daily risk of *Clostridioides difficile* infection at two large academic health centers

Meghana Kamineni, BS,
Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA

Erkin Ötleş, MEng,
Medical Scientist Training Program, University of Michigan Medical School, Ann Arbor, MI

Department of Industrial and Operations Engineering, University of Michigan College of Engineering, Ann Arbor, MI

Jeeheh Oh, PhD,
Division of Computer Science and Engineering, University of Michigan College of Engineering, Ann Arbor, MI

Krishna Rao, MD, MS,
Department of Internal Medicine, Division of Infectious Diseases, University of Michigan Medical School, Ann Arbor, MI

Vincent B. Young, MD, PhD,
Department of Internal Medicine, Division of Infectious Diseases, University of Michigan Medical School, Ann Arbor, MI

Benjamin Y. Li, BS,
Medical Scientist Training Program, University of Michigan Medical School, Ann Arbor, MI

Division of Computer Science and Engineering, University of Michigan College of Engineering, Ann Arbor, MI

Lauren R. West, MPH,
Infection Control Unit, Massachusetts General Hospital, Boston, MA

David C. Hooper, MD,
Infection Control Unit, Massachusetts General Hospital, Boston, MA

Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA

Harvard Medical School, Boston, MA

Erica S. Shenoy, MD, PhD,
Infection Control Unit, Massachusetts General Hospital, Boston, MA

Corresponding Author: Meghana Kamineni, 77 Massachusetts Ave, Cambridge, MA 02142, Tel: (815) 793-1713, kamineni@alum.mit.edu.

Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA

Harvard Medical School, Boston, MA

John G. Guttag, PhD,
Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA

Jenna Wiens, PhD,
Division of Computer Science and Engineering, University of Michigan College of Engineering, Ann Arbor, MI

Maggie Makar, PhD
Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA

Abstract

Many data-driven patient risk stratification models have not been evaluated prospectively. We performed and compared the prospective and retrospective evaluations of two *Clostridioides difficile* infection (CDI) risk-prediction models at two large academic health centers and discuss the models' robustness to dataset shifts.

Keywords

INTRODUCTION

Many data-driven risk prediction models offering the promise of improved patient outcomes have been evaluated retrospectively, but few have been evaluated prospectively.[1–4] Models that are not evaluated prospectively are susceptible to degraded performance because of dataset shifts.[5] Shifts in data can arise from changes in patient populations, hospital procedures, care delivery approaches, epidemiology, and information technology (IT) infrastructure. [2,6]

In this work, we prospectively evaluated a data-driven approach for *Clostridioides difficile* infection (CDI) risk prediction that had previously been shown to achieve high performance in retrospective evaluations at two large academic health centers.[4] This approach models the likelihood of acquiring CDI as a function of patient characteristics. However, this evaluation occurred on retrospective data, and prospective validation is necessary because other models that have not been prospectively evaluated often performed worse when deployed.[7] Risk predictions can guide clinical interventions, including antibiotic de-escalation and duration, beta-lactam allergy evaluation, and isolation.[8]

Using this approach, we trained models for both institutions on initial retrospective cohorts and performed evaluations on retrospective and prospective cohorts. We compared the prospective performance of these models to their retrospective evaluations to determine their

robustness with respect to dataset shifts. By showcasing the robustness of this approach, we provide more support to utilize this approach in clinical workflows.

METHODS

The study included retrospective and prospective periods for adult inpatient admissions to Massachusetts General Hospital (MGH) and Michigan Medicine (MM).

As previously described,[9] patient demographics, admission details, patient history, daily hospitalization information, and exposure and susceptibility to the pathogen (e.g. antibiotic therapy) were extracted from the electronic health record (EHR) of each institution and pre-processed. To consider hospital-onset CDI, we excluded patients who tested positive in the first two calendar days of their admission, stayed less than 3 days, or tested positive in the 14 days before admission. Testing protocols are described in the supplement. A data-driven model to predict risk of hospital-onset CDI was developed for each institution. Each model was based on regularized logistic regression and included 799 and 8,070 variables at MGH and MM, respectively. More aggressive feature selection was applied at MGH to prioritize computational efficiency.[9] For the retrospective evaluation, data were extracted from May 5, 2019 to October 31, 2019 at MGH and July 1, 2019 to June 30, 2020 at MM. For the prospective evaluation, we generated daily extracts of information for all adult inpatients from May 5, 2021 to October 31, 2021 at MGH and July 1, 2020 to June 30, 2021 at MM, keeping the months consistent across validation time periods. We used different periods at the two institutions because of differences in data availability.

When applied to retrospective and prospective data at each institution, the models generated a daily risk score for each patient. We evaluated the discriminative performance of each model at the encounter level using the area under the receiver operator characteristic curve (AUROC). Using thresholds based on the 95th percentile of the retrospective training cohort, we measured the sensitivity, specificity, and positive predictive value (PPV) for each model. 95% confidence intervals were computed using 1,000 Monte-Carlo case-resampled bootstraps. We compared the models' retrospective and prospective performances to understand the impact of any dataset shifts.

This study was approved by the institutional review boards of both participating sites (University of Michigan, Michigan Medicine HUM00147185 and HUM00100254, Mass General Brigham 2012P002359), with a waiver of informed consent.

RESULTS

After applying exclusion criteria, the final retrospective cohort included 18,030 admissions (138 CDI cases) and 25,341 admissions (158 CDI cases) at MGH and MM, respectively. The prospective cohort included 13,712 admissions (119 CDI cases) and 26,864 admissions (190 CDI cases) at MGH and MM, respectively. The demographic characteristics of the study populations are provided (Supplement, Table 1).

At MGH, the model achieved an AUROC of 0.744 (95% confidence interval [CI], 0.707 – 0.781) and 0.748 (95% CI, 0.707–0.791) in retrospective and prospective cohorts,

respectively. At MM, the model achieved an AUROC of 0.778 (95% CI, 0.744 – 0.814) and 0.767 (95% CI, 0.737 – 0.801) in retrospective and prospective cohorts, respectively. The AUROCs for predicting CDI risk on both retrospective and prospective cohorts were similar each month and did not exhibit significant monthly variation throughout either evaluation period (Figure 1). At MGH, the classifiers' sensitivity, specificity, and PPV were 0.138, 0.951, and 0.021 on the retrospective data and 0.210, 0.949, and 0.035 on the prospective data. At MM, the classifiers' sensitivity, specificity, and PPV were 0.215, 0.964, and 0.036 on the retrospective data and 0.189, 0.950, and 0.026 on the prospective data (Figure 2).

DISCUSSION

We evaluated two data-driven institution-specific CDI risk prediction models on prospective cohorts, demonstrating how the models would perform if applied in real-time; that is, how, if implemented with daily data extracts, the models would perform generating daily risk predictions for adult inpatients. The models at both MGH and MM were robust to dataset shift. Notably, the prospective cohorts included patients admitted during the COVID-19 pandemic, whereas the retrospective cohorts did not. Surges in hospital admissions and staff shortages throughout the pandemic affected patient populations and hospital procedures related to infection control; the consistent performance of the models during the COVID-19 pandemic increases confidence that the models are likely to perform well when integrated into clinical workflows. Clinicians can utilize risk predictions to guide interventions, such as isolation and modifying antibiotic administration, and allot limited resources for patients most at risk.[8] These models should be applied to patients meeting the inclusion criteria, and application to a broader cohort may impact results.

Since implementing this methodology requires significant IT support, initial deployment is likely to occur through larger hospitals or EHR vendors, a common approach for risk prediction models.[7] While the methodology is complex, it is handled by the back-end. The interface with clinicians can be quite simple, as the end-user only receives a prediction for each patient.

The PPV was calculated using a threshold based on the 95th percentile of retrospective cohorts. The PPV is between 2.625 and 6 times higher than the pre-test probability, an appropriate level for some interventions, such as beta-lactam allergy evaluations. For interventions requiring higher PPVs, higher thresholds should be used.

Despite the importance of evaluating models prior to deployment, models are rarely prospectively or externally validated.[1–4] Other prior retrospective external validation attempts of models for incident CDI did not replicate the original performance.[10] When performed, prospective and external validation can highlight model shortcomings before integration into clinical workflows. For instance, an external retrospective validation of a widely utilized sepsis prediction model showed that the computed scores at a new institution differed significantly from the model developer's reported validation performance.[7] This model was not tailored to specific institutions, but such discrepancies may still arise with institution-specific models. Especially, when there are many covariates, models can overfit to training data and are therefore susceptible to dataset shifts. In our case, the differences

between retrospective and prospective performances of both models in terms of AUROC were small with large overlapping confidence intervals.

While the successful prospective performance of two institution-specific CDI risk prediction models is encouraging, it does not guarantee that the models will perform well in the face of future dataset shifts. Epidemiology, hospital populations, workflows, and IT infrastructure are constantly changing, thus underscoring that deployed models should be carefully monitored for performance over time.[11]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST

EÖ: patent pending for the University of Michigan for an artificial intelligence-based approach for the dynamic prediction of health states for patients with occupational injuries. KR: Dr. Rao is supported in part from an investigator-initiated grant from Merck & Co, Inc.; he has consulted for Bio-K+ International, Inc., Roche Molecular Systems, Inc., Seres Therapeutics, Inc. and Summit Therapeutics, Inc.

REFERENCES

1. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17:195. [PubMed: 31665002]

2. Brajer N, Cozzi B, Gao M, et al. Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. JAMA Netw Open 2020;3:e1920733.

3. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46:383–400. [PubMed: 31965266]

4. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689. [PubMed: 32213531]

5. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med 2021;385:283–286. [PubMed: 34260843]

6. Menon A, Perry DA, Motyka J, et al. Changes in the Association Between Diagnostic Testing Method, Polymerase Chain Reaction Ribotype, and Clinical Outcomes From Clostridioides difficile Infection: One Institution's Experience. Clinical Infectious Diseases 2021;73:e2883–e2889. [PubMed: 32930705]

7. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med 2021;181:1065–1070. [PubMed: 34152373]

8. Dubberke ER, Carling P, Carrico R, et al. Strategies to prevent Clostridium difficile infections in acute care hospitals: 2014 update. Infect Control Hosp Epidemiol 2014;35:628–645. [PubMed: 24799639]

9. Oh J, Makar M, Fusco C, et al. A Generalizable, Data-Driven Approach to Predict Daily Risk of Clostridium difficile Infection at Two Large Academic Health Centers. Infect Control Hosp Epidemiol 2018;39:425–433. [PubMed: 29576042]

10. Perry DA, Shirley D, Micic D, et al. External Validation and Comparison of Clostridioides difficile Severity Scoring Systems. Clin Infect Dis 2021.

11. Otles E, Oh J, Li B, et al. Mind the Performance Gap: Examining Dataset Shift During Prospective Validation. Proceedings of the 6th Machine Learning for Healthcare Conference; 2021; Proceedings of Machine Learning Research.
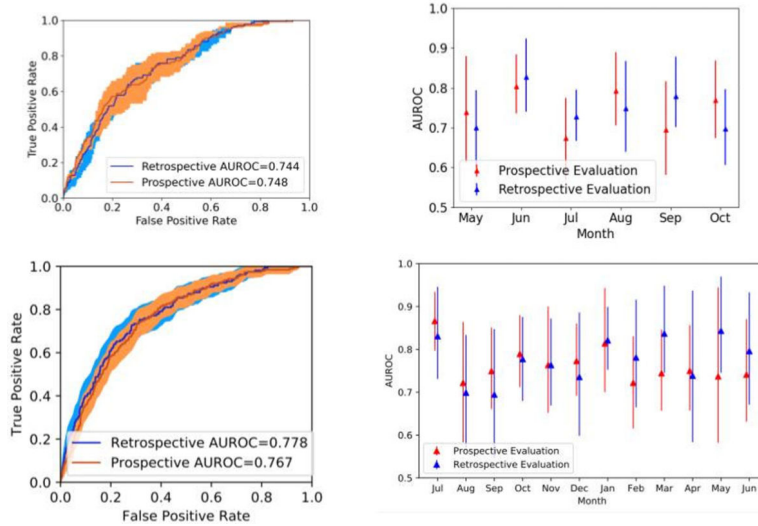
Figure 1: AUROC at MGH and MM in Retrospective and Prospective Evaluations.
The figures on the left show a comparison of AUROC in retrospective and prospective evaluations MGH (upper) and MM (lower). The 95% confidence intervals (CI) for the AUROC are shaded. The figures on the right show a monthly AUROC comparison at MGH (upper) and MM (lower). The 95% CI for the AUROC are represented by error bars.
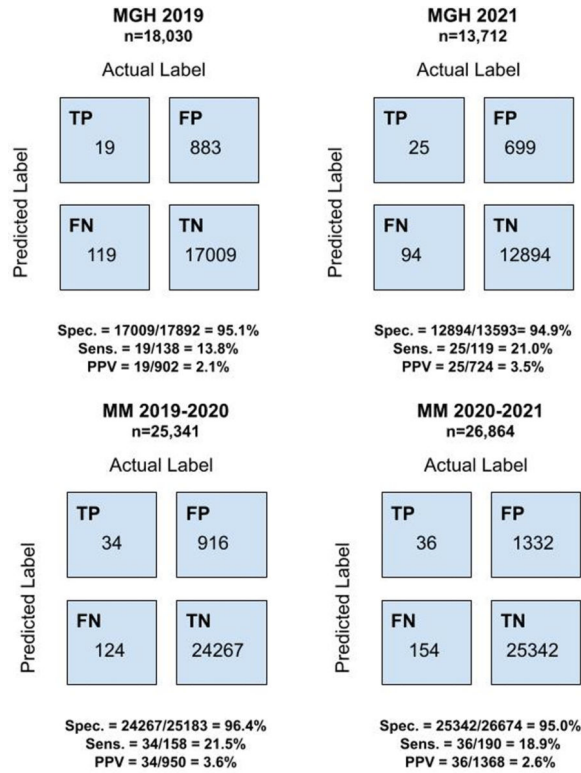
Figure 2: Confusion Matrices at MGH and MM in Retrospective and Prospective Evaluations.
The figures on the left display confusion matrices, sensitivity, specificity, and positive predictive values for retrospective evaluations at MGH (upper) and MM (lower). The figures on the right display the same metrics for prospective evaluations at MGH (upper) and MM (lower).