Open Forum Infectious Diseases

MAJOR ARTICLE

Machine Learning in Infectious Disease for Risk Factor Identification and Hypothesis Generation: Proof of Concept Using Invasive Candidiasis

Lisa M. Mayer,[1] Jeffrey R. Strich,[2] Sameer S. Kadri,[2] Michail S. Lionakis,[3] Nicholas G. Evans,[4] D. Rebecca Prevots,[5] and Emily E. Ricotta[5]

[1]Office of Data Science and Emerging Technologies, Office of Science Management and Operations, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Rockville, Maryland, USA, [2]Critical Care Medicine Department, NIH Clinical Center, NIH, Bethesda, Maryland, USA, [3]Fungal Pathogenesis Section, Laboratory of Clinical Immunology & Microbiology (LCIM), NIAID, NIH, Bethesda, Maryland, USA, [4]Department of Philosophy, University of Massachusetts Lowell, Lowell, Maryland, USA, and [5]Epidemiology and Population Studies Unit, LCIM, NIAID, NIH, Bethesda, Maryland, USA

Background. Machine learning (ML) models can handle large data sets without assuming underlying relationships and can be useful for evaluating disease characteristics, yet they are more commonly used for predicting individual disease risk than for identifying factors at the population level. We offer a proof of concept applying random forest (RF) algorithms to *Candida*-positive hospital encounters in an electronic health record database of patients in the United States.

Methods. *Candida*-positive encounters were extracted from the Cerner *HealthFacts* database; invasive infections were laboratory-positive sterile site *Candida* infections. Features included demographics, admission source, care setting, physician specialty, diagnostic and procedure codes, and medications received before the first positive *Candida* culture. We used RF to assess risk factors for 3 outcomes: any invasive candidiasis (IC) vs non-IC, within-species IC vs non-IC (eg, invasive *C. glabrata* vs noninvasive *C. glabrata*), and between-species IC (eg, invasive *C. glabrata* vs all other IC).

Results. Fourteen of 169 (8%) variables were consistently identified as important features in the ML models. When evaluating within-species IC, for example, invasive *C. glabrata* vs non-invasive *C. glabrata*, we identified known features like central venous catheters, intensive care unit stay, and gastrointestinal operations. In contrast, important variables for invasive *C. glabrata* vs all other IC included renal disease and medications like diabetes therapeutics, cholesterol medications, and antiarrhythmics.

Conclusions. Known and novel risk factors for IC were identified using ML, demonstrating the hypothesis-generating utility of this approach for infectious disease conditions about which less is known, specifically at the species level or for rarer diseases.

Keywords. artificial intelligence; big data; infectious diseases; invasive candidiasis; machine learning.

In recent years, novel methods using artificial intelligence (AI) and machine learning (ML) have been developed and applied to different infectious diseases to enhance clinician decision-making [1, 2], an application broadly called "clinical decision support" [3]. They have been used to predict the onset of outbreaks [4, 5], assign risk scores to individual patients for clinical outcomes [6, 7], and identify clinically predictive biomarkers of disease progression [8, 9].

While ML models are increasingly popular in clinical contexts, they may lack utility or fail in their goal for reasons including bias in the training data (eg, all the patients are Caucasian, trying to predict for non-Caucasian patients), nonstationarity of the outcome (eg, prediction of antibiotic resistance leads to improved stewardship, which results in changing resistance patterns), changes in data structures (eg, antibiotic resistance threshold definitions change over time), and a lack of robust diagnostic criteria (eg, low diagnostic sensitivity). Models aimed at predicting individual patient outcomes may also, if not adequately validated, do so based on spurious factors: Famously, an algorithm predicted pneumonia based on the brand of computed tomography scanners used in an image [10].

ML models for the identification of population-level risk factors and hypothesis generation are less well explored in the infectious disease space. In classic (non-ML-based) regression models, variables believed to be population-level risk factors for a disease are chosen a priori, requiring knowledge and assumptions about diseases and their characteristics, or through statistical methods like stepwise selection that rely on statistical significance. The limitation of these methods is that they can introduce bias into the analysis due to prior conceptions about risk, in the case of a priori selection, and can result in the selection of spuriously related variables and exclude real explanatory variables when using stepwise methods [11]. In both cases, since variables must be predetermined by the research team, unknown risk factors can be excluded entirely from the

Received 28 July 2022; editorial decision 30 July 2022; accepted 02 August 2022; published online 3 August 2022

Correspondence: E. E. Ricotta, PhD, Epidemiology and Population Studies Unit, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane 7D18, Rockville, MD 20852 (emily.ricotta@nih.gov).

Open Forum Infectious Diseases®

https://doi.org/10.1093/ofid/ofac401

analysis. The advantage of using ML models is their ability to handle a large amount of structured data without assuming underlying relationships, as well as not requiring deliberate specification of a relationship between variables (ie, interaction terms), allowing a more flexible and potentially less biased approach to evaluating characteristics of patients with diseases. Appropriate design of these algorithms allows the incorporation of multiple data streams including lab values, demographics, medications, and comorbidities into a single algorithm. This application of ML has been used in identifying risk factors for health care–associated meningitis in Russia [12] and Rift Valley Fever outbreaks in Africa [13] but has been limited to clinical prediction models of individual patient risk in the fungal disease literature [14–18].

In this paper, we provide a proof of concept for the utility of ML for identifying population-level risk factors of infectious diseases, focusing specifically on evaluating the differences in risk factors of invasive candidiasis (IC) compared with non-IC hospitalizations, as well as between and within *Candida* species. To do this, we applied random forest (RF) algorithms to *Candida*-positive hospital encounters in an electronic health record database of patients from all regions of the United States. IC is one of the most important hospital-associated infections in the United States and requires early diagnosis to ensure the correct therapy [19]; however, the lack of quality diagnostics makes it difficult for ML algorithms to predict which individual patients might develop IC, especially without a robust understanding of species-level risk factors of invasive disease. We demonstrate that our algorithms identify both known and unknown risk factors for IC, helping to generate new hypotheses to advance clinical research and assist clinicians with decision-making, especially in the absence of reliable laboratory findings [20]. We then discuss the considerations and challenges for developing these ML models.

METHODS

Ethical Review of Study and Waiver of Consent

The National Institutes of Health Office of Human Research Protections determined this research to be not human subjects research and therefore exempt from institutional review board review, as analyses were limited to existing deidentified data.

Study Population

Data were extracted from the Cerner *HealthFacts* database, a linked electronic health record (EHR) database containing inpatient and outpatient hospital encounter data for >63 million patients throughout the United States [21]. From 2009 to 2017, any encounter with a laboratory-positive *Candida* test result was initially retained for this analysis. Of note, we did not require a negative culture upon admission, so infections could be hospital or community acquired. Encounters with specimens isolated from respiratory or stool samples or specimens containing multiple *Candida* species were excluded. Distinct positive encounters detected ≤2 days apart (eg, an emergency room visit resulting in a hospital admission) were treated as 1 episode of *Candida* infection.

Outcome Assessment

The outcome measure in this analysis was IC, which was assigned based on the body source of the encounter's *Candida*-positive culture. Sterile site infections were determined using a combination of specimen collection source and body site (eg, blood, venous), as well as the specific laboratory procedure used for isolation of the organism (eg, blood culture). The final list of microbiology tests for inclusion was determined by clinical consensus (see Ricotta et al. for complete list [22]). Hospital encounters with laboratory-positive *Candida* infections in a sterile site were identified as invasive infections. *Candida*-positive encounters with isolates from a nonsterile site were considered noninvasive infections.

A secondary outcome was to evaluate species-specific IC, including *C. albicans*, *C. glabrata*, *C. parapsilosis*, *C. tropicalis*, and other *Candida* spp. Risk factors for IC vs non-IC by species were evaluated by comparing characteristics of IC encounters with non-IC encounters within species (ie, invasive *C. albicans* with noninvasive *C. albicans*). Risk factors for species-specific IC were evaluated by comparing characteristics of encounters between *Candida* species among IC encounters only (ie, invasive infections by *C. albicans* vs invasive infections by another *Candida* spp.).

Variable Selection

For each encounter, we extracted variables including patient demographics, medication classes, laboratory and microbiology test results, procedure codes, admission source, care setting, medical specialty of the attending physician, and International Classification of Diseases, 9th and 10th Revision (ICD-9 and ICD-10), codes for comorbidities and health history. To the best of our ability (ie, when dates were available), we restricted the presence of variables to those before the date of the first positive *Candida* culture. Hospital information such as size, geographic location, and teaching status was also included. ICD codes were used to calculate the Elixhauser comorbidity index [23] for every encounter using the R package ICD [24] and were then categorized using the Healthcare Cost and Utilization Project's Clinical Classification Software for ICD-9 and ICD-10, respectively [25, 26]. Medications were classified using RxNorm and FDA National Drug Codes [27, 28]. Laboratory procedures and clinical events were ultimately excluded from the analysis as this resulted in too many overly general features and tracked standard hospital practice rather than IC (eg, the overwhelming majority of laboratory test results were "normal" regardless of test type, IC status, or species).

Model Training and Validation

RF, a decision tree–based machine learning algorithm, was selected due to its ability to handle many variables and estimate their importance in classifying an outcome. Three separate models were evaluated using rand_forest from R tidymodels [29]: invasive vs noninvasive candidiasis among all encounters, within-species invasive vs noninvasive candidiasis, and between-species invasive infections. To assess the relative importance of the variables (risk factors), permutation importance was calculated for each variable in the final models. Permutation importance ranks each variable by how much its presence and value impact the model's ability to classify the observation, for example, as IC vs non-IC, over multiple model iterations. Relative importance is then determined by dividing each variable's permutation importance score by the largest importance score of the variables for each of the 10 iterations, then multiplying by 100. Importantly, these values are not risk estimates and should not be interpreted like odds or risk ratios, nor do they imply directionality of the outcome; rather, a feature with a higher importance score will have a larger impact on the classification of hospitalizations as IC or non-IC [30].

More information on methodology is available in the Supplementary Data. Analysis was done in R, version 4.0.3-4, and RStudio, version 1.3.1056 [31, 32].

RESULTS

Patient demographics are broken down by IC vs non-IC (Supplementary Table 1). After data cleaning, we were left with 169 features for analysis (Supplementary Tables 2 and 3). Models are listed in Supplementary Table 4; model performance is summarized in Supplementary Table 5. Feature distribution for models 2–11 (within-species IC vs non-IC and between-species IC) is shown in Supplementary Table 6, which informs the directionality of the association between variable and outcome. Illustrative examples of each outcome using *C. glabrata* are provided below, with complete results available in the Supplementary Data.

Population Characteristics

Between 2009 and 2017, there were 19 381 763 unique inpatient encounters identified at 203 hospitals that reported ≥1 case of *Candida* in their facility over this period. Of these, 172 120 were microbiologically confirmed *Candida*-positive inpatient hospitalizations. After excluding encounters where *Candida* was isolated only from the stool or a respiratory source, where multiple *Candida* species were isolated, or where the *Candida* species was not reported, we had 116 725 unique encounters for analysis, of which 14 311 (12%) were classified as invasive infections. Species distribution was 61% *C. albicans* (11% invasive), 13% *C. glabrata* (21% invasive), 4% *C. parapsilosis* (31% invasive), 5% *C. tropicalis* (16% invasive), and 17% other *Candida* spp. (6% invasive).

Overall IC vs Non-IC

In this model, 14 variables were consistently ranked in the top 10 most important features across all iterations (Figure 1). Variables found in a higher proportion of IC-positive hospitalizations were having a central venous catheter (CVC), undergoing an operation on the digestive system, receiving hemostasis modifiers, receipt of parenteral nutrition (TPN), infection with *C. glabrata* or *C. parapsilosis*, being in the intensive care unit (ICU) or having an unknown care setting, and male sex. IC-positive hospitalizations also had more time from admission to *Candida* culture positivity and a higher Elixhauser comorbidity index. Those found in a higher proportion of non-IC hospitalizations were being age 65+ years, being seen in the emergency room (ER), and being female (Supplementary Tables 3 and 6).

Within-Species IC vs Non-IC: *C. glabrata* IC vs *C. glabrata* Non-IC Infections

Invasive infections with *C. glabrata* were generally seen more often in female individuals who were hospitalized in urban and teaching hospitals than those infected with noninvasive *C. glabrata*. These individuals had more ICU visits, CVCs, and TPN receipt and were less likely to be seen in the ER. Patients were more likely to have received anesthetics and gastrointestinal agents and had both more operations on the digestive system and more miscellaneous diagnostic and therapeutic procedures. IC hospitalizations had more time to culture positivity and had a slightly higher Elixhauser score than non-IC patients (Figure 2). Results from the remaining within-species IC vs non-IC models can be found in the Supplementary Data.

Between-Species IC: Invasive *C. glabrata* vs Non–*C. glabrata* Invasive Disease

Individuals hospitalized with invasive *C. glabrata* were older, had more chronic renal disease than IC infections with non-*glabrata* *Candida* species, and were found in larger teaching hospitals. These individuals had more ICU visits and received antihyperlipidemic medications, cardiac (antiarrhythmic) agents, diabetes therapy, diuretics, and vascular agents more frequently than non-*glabrata* IC. These patients also had a higher prevalence of a prior positive *C. glabrata* culture. Time from admission to first positive culture and Elixhauser score were the same for both groups but were consistently highly important in this model (Figure 3). Results from the remaining between-species IC models can be found in the Supplementary Data.

DISCUSSION

In this study, we provide a mechanism for identifying risk factors for overall and species-specific IC, both known and novel.

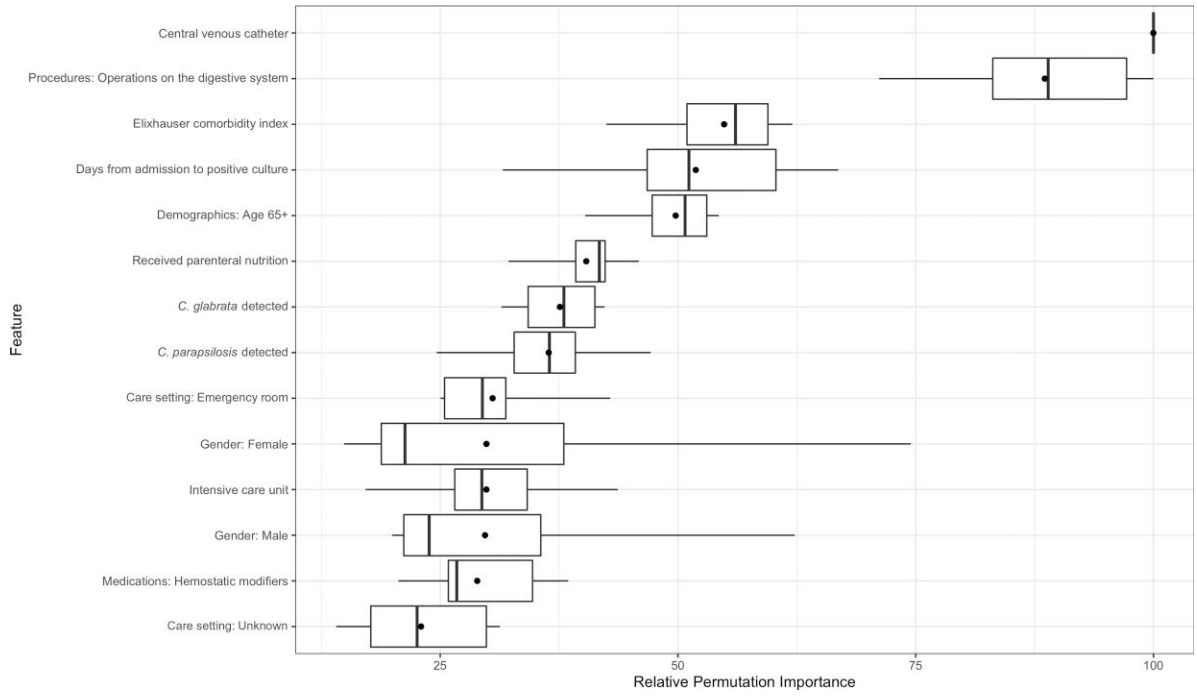

Figure 1. Features appearing in the 10 highest relative permutation importance values for classifying overall invasive vs noninvasive candidiasis. Boxplots are the variable's minimum, first quartile, median, mean (black dot), third quartile, and maximum relative importance values across the 10 model iterations. Importance is ordered by average relative importance across all iterations.

Whereas the identification of known risk factors provides clinical validation and confidence in our approach, the discovery of novel risk factors generates new hypotheses and identifies additional, potentially actionable, factors for this disease. This mechanism of investigation can easily be applied to other organism models to the same effect.

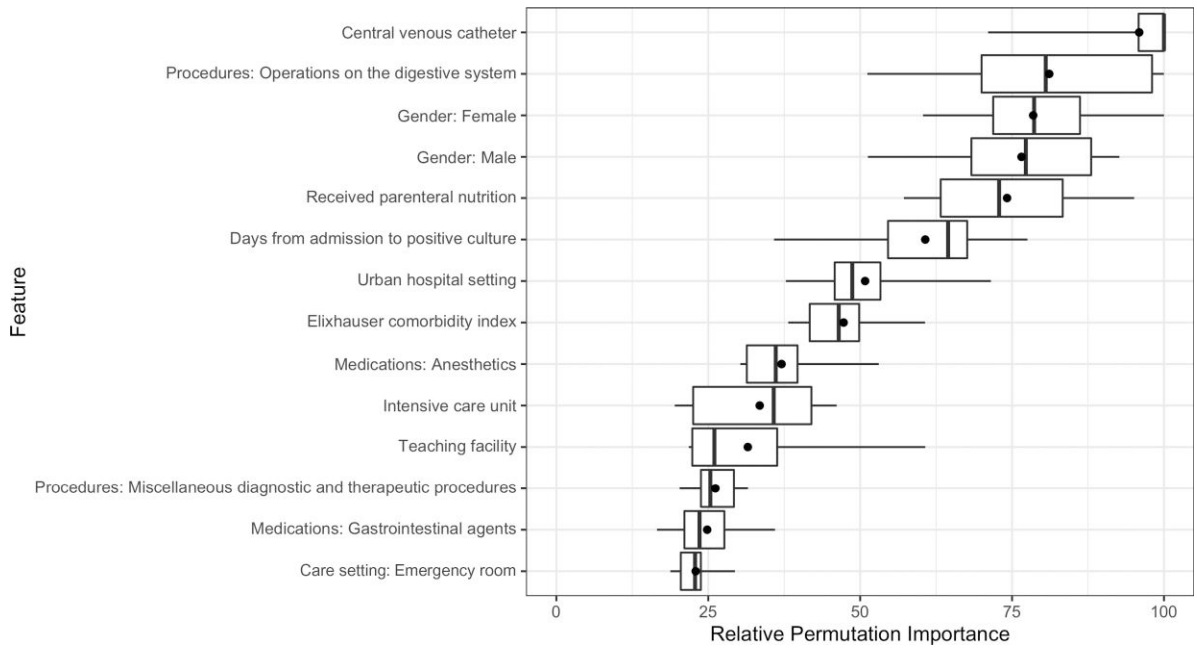

Figure 2. Features appearing in the 10 highest relative permutation importance values for classifying *C. glabrata* IC vs noninvasive *C. glabrata*. Boxplots are the variable's minimum, first quartile, median, mean (black dot), third quartile, and maximum relative importance values across the 10 model iterations. Importance is ordered by average relative importance across all iterations.

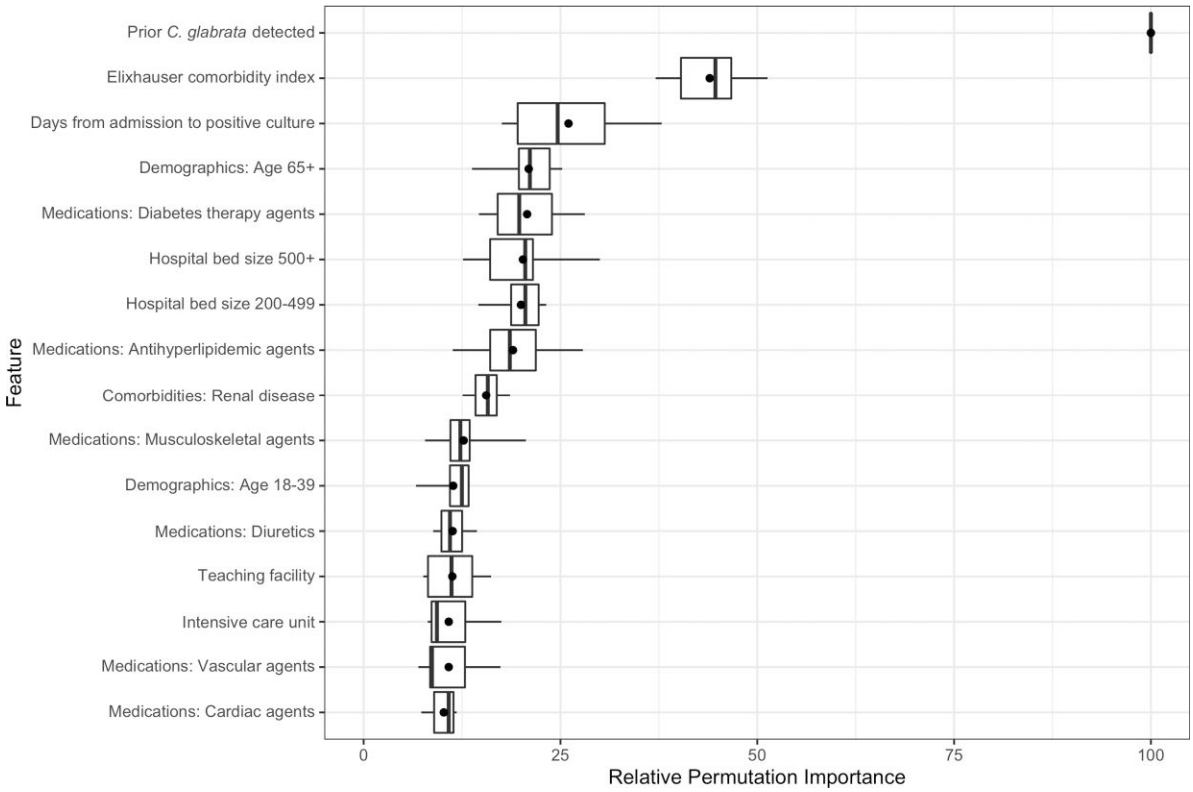

Figure 3. Features appearing in the 10 highest relative permutation importance values for classifying *C. glabrata* IC vs non-*glabrata* IC. Boxplots are the variable's minimum, first quartile, median, mean (black dot), third quartile, and maximum relative importance values across the 10 model iterations. Importance is ordered by average relative importance across all iterations. Abbreviation: IC, invasive candidiasis.

The uniqueness of using a large EHR data set was evident in our ability to assess within-species IC vs non-IC (models 2–6) and between-species IC (models 7–11), as this is typically impossible to do in single-center or small multicenter studies. Determining species-specific risk factors for IC is helpful for several clinical reasons, including providing a better guide for empiric treatment and directing source control strategies. Studies have demonstrated that early detection and treatment of IC with appropriate antifungal medications are important for effective management [33–35]; however, current diagnostic tools suffer from low sensitivity. For example, blood culture, the gold standard for detection of *Candida* infection, has only a 50% sensitivity; sensitivity for infection in deeper sites is even lower [19, 20]. Lack of diagnostic power may delay treatment and cause inappropriate empiric antifungal therapy, especially among *Candida* species known to be antifungal-resistant, including *C. glabrata* [19, 36]. However, having a sense of species-specific clinical phenotypes can help clinicians determine whether further investigation or empiric treatment could be warranted and can provide more evidence for determining whether the patient might be infected with a species of *Candida* that has higher intrinsic antifungal resistance.

While many of the features we identified across the various models are known risk factors, there are some that are not themselves commonly associated with IC in the literature. In some cases, these do not help identify a *Candida* infection specifically: Anesthetics were identified by the model but are unlikely to have a direct relationship with IC; however, people undergoing surgery will receive anesthetics for the procedure, and surgery is a strong predictor of IC [18, 19]. The same can be said of features such as Elixhauser comorbidity index: IC is typically associated with sicker patients. In other cases, variables not identified as important may themselves provide important insights into IC. For example, *C. tropicalis* has been associated with malignancy in the literature [37]; however, we did not find malignancy to be a variable of importance for classifying invasive *C. tropicalis* either compared with noninvasive *C. tropicalis* or with IC caused by other *Candida*. Findings like this can challenge the status quo in diagnosing and classifying IC and highlight areas for further research. This emphasizes the utility of ML as a tool for risk factor analysis as the identification of features that are uncommonly thought of as important for IC or are overlooked entirely can lead to new avenues for clinical exploration in the field. The relevance of particular features for classifying a hospitalization as

IC vs non-IC might not be straightforward or obvious, necessitating thoughtful and open-minded discussion of all features selected by the model including those that are of known importance, those that are of previously unknown importance, and those considered historically "nonimportant."

Two examples of this in our analysis are renal disease and associated medications in invasive *C. glabrata* vs other IC and features representing cardiac disorders across models. While it is known that individuals with renal disease are generally sicker and require more invasive procedures such as catheterization and peritoneal dialysis, both known risk factors for IC, there is little to no evidence in the literature associating *C. glabrata* specifically with this condition or its treatments. This association warrants further investigation, as this patient group could be unknowingly exposed to increased risk of *C. glabrata* infections through previously unknown or unrecognized mechanisms.

The second example of cardiac disorders is also uncommonly found in the literature to strongly support an association (positive or negative) between these illnesses and IC of any species; however, evidence exists of an interaction between invasive fungal infections and platelets (importance in this study indicated by the comorbidity coagulopathy [38]), notably through the observation of the presence of thrombocytopenia in candidemic individuals and the fact that platelets themselves play a role in anti-*Candida* immunity [39, 40]. Another example is that IC can be associated with the development of endocarditis [41], which can be comorbid with congestive heart failure [42] and cardiac operations like prosthetic valve placement [43]. These results from our analysis likely do not represent a causal association between these features and IC but rather indicate increased cardiac involvement in these types of infection. For example, it is possible that, while we included medications only if they were dispensed before the first positive *Candida* culture, that first culture does not accurately reflect the time when the infection started due to the aforementioned low sensitivity of microbiological identification methods. This can result in misdiagnoses, which lead to delayed prescribing of antifungal drugs and increases the likelihood of cardiac involvement and death. Additionally, because cardiac procedures such as operations on the cardiovascular system are identified using ICD-9/10 codes, their timing relative to IC incidence is unknown and could occur after positivity, which has been seen in the literature [44, 45].

Another strength of ML algorithms is that they are easily able to model nonlinear relationships between variables and the outcome as well between the predictor variables themselves. This does require careful interpretation of important features and their timing in relation to the event of interest to ensure that valid conclusions are drawn from the model. Subject matter experts like clinicians and laboratory scientists should be involved in the analytic process to interpret variables identified by the model and ensure appropriateness of any data manipulation (eg, combining categories, ensuring biologically relevant timing of variables). Epidemiologists and data scientists are required to oversee data usage and algorithm selection and to contextualize model output. Finally, studies should be explicit about how features are chosen for inclusion in algorithms, including whether they are being selected a priori based on current knowledge or from first principles.

It is also useful to consider features that are deliberately *excluded* from the model, in this case, laboratory procedures. We began with laboratory procedures in the model but found that they generated noise that overpowered the algorithm, resulting in relatively uninformative features. Inclusion of raw laboratory values would require either extensive reduction and modification by subject matter experts (contradicting the ejective of this methodology), or some alternative variable selection step before the RF analysis. It is possible that as new methods are developed in the field there will be better ways to utilize these data to obtain informative information; however, methods development was outside the scope of this paper. We therefore opted to remove those features entirely, which also provided the benefit of requiring only features that are common to any clinician's EHR.

These models require maintenance to ensure that the outputs of the model remain valid over time. The multidisciplinary group suggested above will be required to continuously ensure that features in the model are still appropriate based on changing scientific knowledge, that changes in clinical measurements and interpretation are accurately captured in the model, and that information gleaned from the output is interpreted in the proper context and with due considerations of the limitations and biases of the algorithm. This is consistent with recent calls for ML models used in clinical practice to undergo continuous risk monitoring [46] and may build capacity for a particular model (or that model in a particular health care setting) to be maintained to a high standard of safety and reliability. However, with proper planning and guidance, ML can be a powerful tool for understanding population-level risk factors for infectious diseases. Indeed, for IC, future analyses might use other clinically relevant distinctions such as comparing bloodstream vs other types of IC, as candidemia and abdominal candidiasis are overlapping but unique diseases with shared and distinct risk factors. The methods presented in this analysis can be readily applied to this and other infectious diseases.

CONCLUSIONS

ML can be used to identify risk factors for infectious diseases, offering an opportunity to study unknown host associations that might inform the epidemiology of and susceptibility to rare and novel pathogens, even among species and subtypes that are typically too few in number to assess in standard clinical studies. These methods provide benefits for building algorithms using multiple data streams that include unstructured,

nonlinear, and correlated clinical data, and while their construction and maintenance must be approached in a multidisciplinary manner, their use can enhance discovery of patient and pathogen factors, thus supporting infectious disease prevention and clinical practice.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

This study used the Office of Cyber Infrastructure and Computational Biology (OCICB) High Performance Computing (HPC) cluster at the National Institute of Allergy and Infectious Diseases (NIAID; Bethesda, MD, USA).

Financial support. This work was supported in part by the Division of Intramural Research of the National Institute of Allergy and Infectious Diseases and the National Institutes of Health Clinical Center. L.M.M. was supported by an appointment to the National Institute of Allergy and Infectious Diseases (NIAID) Emerging Leaders in Data Science Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy (DOE) and NIAID. ORISE is managed by ORAU under DOE contract number DE-SC0014664. N.G.E. is supported by the National Science Foundation (1734521), the Greenwall Foundation Faculty Scholars Program, the Davis Educational Foundation, and the US Air Force Office of Scientific Research (FA9550-21-1-0142).

Potential conflicts of interest. The authors declare no conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

Author contributions. The authors confirm contributions to the paper as follows: study conception and design: E.E.R., D.R.P.; data acquisition: S.S.K.; data analysis: L.M.M., E.E.R.; interpretation of results: L.M.M., J.R.S., E.E.R.; draft manuscript preparation: L.M.M., N.G.E., E.E.R. All authors reviewed the results and approved the final version of the manuscript.

Data availability. Data not publicly available.

References

1. Roth JA, Battegay M, Juchler F, Vogt JE, Widmer AF. Introduction to machine learning in digital healthcare epidemiology. Infect Control Hosp Epidemiol **2018**; 39:1457–62.
2. Peiffer-Smadja N, Rawson TM, Ahmad R, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect **2020**; 26:584–95.
3. HealthIT.gov. Clinical decision support. Available at: https://www.healthit.gov/topic/safety/clinical-decision-support. Accessed May 18, 2022.
4. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinf **2014**; 15:276.
5. Chae S, Kwon S, Lee D. Predicting infectious disease using deep learning and big data. Int J Environ Res Public Health **2018**; 15:1596.
6. Chiu HYR, Hwang CK, Chen SY, et al. Machine learning for emerging infectious disease field responses. Sci Rep **2022**; 12:328.
7. Luz CF, Vollmer M, Decruyenaere J, Nijsten MW, Glasner C, Sinha B. Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies. Clin Microbiol Infect **2020**; 26:1291–9.
8. Lamping F, Jack T, Rübsamen N, et al. Development and validation of a diagnostic model for early differentiation of sepsis and non-infectious SIRS in critically ill children—a data-driven approach using machine-learning algorithms. BMC Pediatr **2018**; 18:112.
9. Taneja I, Reddy B, Damhorst G, et al. Combining biomarkers with EMR data to identify patients in different phases of sepsis. Sci Rep **2017**; 7:10800.
10. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med **2018**; 15:e1002683.
11. Smith G. Step away from stepwise. J Big Data **2018**; 5:32.
12. Savin I, Ershova K, Kurdyumova N, et al. Healthcare-associated ventriculitis and meningitis in a neuro-ICU: incidence and risk factors selected by machine learning approach. J Crit Care **2018**; 45:95–104.
13. Walsh MG, Willem de Smalen A, Mor SM. Wetlands, wild *Bovidae* species richness and sheep density delineate risk of rift valley fever outbreaks in the African continent and Arabian Peninsula. PLoS Negl Trop Dis **2017**; 11:e0005756.
14. Hermsen ED, Zapapas MK, Maiefski M, Rupp ME, Freifeld AG, Kalil AC. Validation and comparison of clinical prediction rules for invasive candidiasis in intensive care unit patients: a matched case-control study. Crit Care **2011**; 15:R198.
15. Shahin J, Allen EJ, Patel K, et al. Predicting invasive fungal disease due to *Candida* species in non-neutropenic, critically ill, adult patients in United Kingdom critical care units. BMC Infect Dis **2016**; 16:480.
16. Playford EG, Lipman J, Jones M, et al. Problematic dichotomization of risk for intensive care unit (ICU)–acquired invasive candidiasis: results using a risk-predictive model to categorize 3 levels of risk from a multicenter prospective cohort of Australian ICU patients. Clin Infect Dis **2016**; 63:1463–9.
17. Guillamet CV, Vazquez R, Micek ST, Ursu O, Kollef M. Development and validation of a clinical prediction rule for candidemia in hospitalized patients with severe sepsis and septic shock. J Crit Care **2015**; 30:715–20.
18. Rauseo AM, Aljorayid A, Olsen MA, et al. Clinical predictive models of invasive *Candida* infection: a systematic literature review. Med Mycol **2021**; 59:1053–67.
19. McCarty TP, White CM, Pappas PG. Candidemia and invasive candidiasis. Infect Dis Clin North Am **2021**; 35:389–413.
20. Clancy CJ, Nguyen MH. Diagnosing invasive candidiasis. J Clin Microbiol **2018**; 56:e01909–17.
21. DeShazo JP, Hoffman MA. A comparison of a multistate inpatient EHR database to the HCUP nationwide inpatient sample. BMC Health Serv Res **2015**; 15:384.
22. Ricotta EE, Lai YL, Babiker A, et al. Invasive candidiasis species distribution and trends, United States, 2009–2017. J Infect Dis **2021**; 223:1295–302.
23. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care **2005**; 43:1130–9.
24. Wasey JO. ICD: comorbidity calculations and tools for ICD-9 and ICD-10 codes (package version 4.0.9.9000). **2018**. Available at: https://cran.r-project.org/package=icd. Accessed March 2021.
25. Healthcare Cost and Utilization Project (HCUP). Clinical Classifications Software (CCS) for ICD-9-CM. **2017**. Available at: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed July 20, 2021.
26. Healthcare Cost and Utilization Project (HCUP). Clinical classifications software refined (CCSR) for ICD-10-PCS procedures. **2021**. Available at: https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp. Accessed July 20, 2021.
27. RxNorm. Available at: https://www.nlm.nih.gov/research/umls/rxnorm/index.html. Accessed July 26, 2022.
28. Food and Drug Administration. openFDA. Available at: https://open.fda.gov/data/ndc/. Accessed July 26, 2022.
29. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. **2020**. Available at: https://www.tidymodels.org. Accessed December 2020.
30. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett **2010**; 31:2225–36.
31. R Core Team. R: a language and environment for statistical computing (version 4.0.3 and version 4.0.4). **2021**. Available at: https://www.R-project.org/.
32. RStudio Team. RStudio: integrated development environment for R (version 1.3.1056). **2020**. Available at: http://www.rstudio.com/.
33. Hsu DI, Nguyen M, Nguyen L, Law A, Wong-Beringer A. A multicentre study to evaluate the impact of timing of caspofungin administration on outcomes of invasive candidiasis in non-immunocompromised adult patients. J Antimicrob Chemother **2010**; 65:1765–70.
34. Morrell M, Fraser VJ, Kollef MH. Delaying the empiric treatment of *Candida* bloodstream infection until positive blood culture results are obtained: a potential risk factor for hospital mortality. Antimicrob Agents Chemother **2005**; 49:3640–5.
35. Arnold HM, Micek ST, Shorr AF, et al. Hospital resource utilization and costs of inappropriate treatment of candidemia. Pharmacother **2010**; 30:361–8.
36. Hadrich I, Ayadi A. Epidemiology of antifungal susceptibility: review of literature. J Mycol Med **2018**; 28:574–84.
37. Wingard JR. Importance of *Candida* species other than *C. albicans* as pathogens in oncology patients. Clin Infect Dis **1995**; 20:115–25.

38. Elixhauser comorbidity—coagulopathy. Available at: https://phenotypes.mpog.org/Elixhauser%20Comorbidity%20-%20Coagulopathy. Accessed January 3, 2022.
39. Netea MG, Joosten LAB, van der Meer JW, Kullberg BJ, van de Veerdonk FL. Immune defence against *Candida* fungal infections. Nat Rev Immunol **2015**; 15:630–42.
40. Eberl C, Speth C, Jacobsen ID, et al. *Candida*: platelet interaction and platelet activity in vitro. J Innate Immun **2019**; 11:52–62.
41. Ioannou P, Volosyraki M, Mavrikaki V, et al. *Candida parapsilosis* endocarditis. Report of cases and review of the literature. Germs **2020**; 10:254–9.
42. Elixhauser comorbidity—congestive heart failure. Available at: https://phenotypes.mpog.org/Elixhauser%20Comorbidity%20-%20Congestive%20Heart%20Failure. Accessed January 3, 2022.
43. Arnold CJ, Johnson M, Bayer AS, et al. *Candida* infective endocarditis: an observational cohort study with a focus on therapy. Antimicrob Agents Chemother **2015**; 59:2365–73.
44. Giacobbe DR, Salsano A, Del Puente F, et al. Risk factors for candidemia after open heart surgery: results from a multicenter case-control study. Open Forum Infect Dis **2020**; 7:XXX–XX.
45. Pasero D, De Rosa FG, Rana NK, et al. Candidemia after cardiac surgery in the intensive care unit: an observational study. Interact Cardiovasc Thorac Surg **2011**; 12:374–8.
46. Babic B, Gerke S, Evgeniou T, Cohen IG. Algorithms on regulatory lockdown in medicine. Science **2019**; 366:1202–4.