American Journal of Infection Control ■■ (2018) ■■-■■

Contents lists available at ScienceDirect

American Journal of Infection Control

journal homepage: www.ajicjournal.org

Major Article

Assessing patient risk of central line-associated bacteremia via machine learning

Cole Beeler MD [a,*], Lana Dbeibo MD [a], Kristen Kelley MPH, CIC [b], Levi Thatcher PhD [c], Douglas Webb MD [b], Amadou Bah MS [b], Patrick Monahan PhD [d], Nicole R. Fowler PhD, MHSA [e], Spencer Nicol [c], Alisa Judy-Malcolm RN, MPH [f], Jose Azar MD [g]

[a] *Indiana University School of Medicine, Indianapolis, IN*
[b] *Infection Prevention for IU Health, Indianapolis, IN*
[c] *Health Catalyst, Salt Lake City, UT*
[d] *Department of Biostatistics, Indiana University, Indianapolis, IN*
[e] *Department of Medicine, Indiana University, Indianapolis, IN*
[f] *IU Health, Indianapolis, IN*
[g] *Indiana University School of Medicine, IU Health, Indianapolis, IN*

Key Words:
Central line-associated bloodstream infection (CLABSI)
Infection control
Machine learning
Infection prediction
Nosocomial infection
Quality improvement

Background: Central line-associated bloodstream infections (CLABSIs) contribute to increased morbidity, length of hospital stay, and cost. Despite progress in understanding the risk factors, there remains a need to accurately predict the risk of CLABSIs and, in real time, prevent them from occurring.
Methods: A predictive model was developed using retrospective data from a large academic healthcare system. Models were developed with machine learning via construction of random forests using validated input variables.
Results: Fifteen variables accounted for the most significant effect on CLABSI prediction based on a retrospective study of 70,218 unique patient encounters between January 1, 2013, and May 31, 2016. The area under the receiver operating characteristic curve for the best-performing model was 0.82 in production.
Discussion: This model has multiple applications for resource allocation for CLABSI prevention, including serving as a tool to target patients at highest risk for potentially cost-effective but otherwise time-limited interventions.
Conclusions: Machine learning can be used to develop accurate models to predict the risk of CLABSI in real time prior to the development of infection.

* Address correspondence to Cole Beeler, MD, Indiana University School of Medicine, Emerson Hall Room 421, 545 Barnhill Dr., Indianapolis, IN 46202.
E-mail address: cbeeler@iupui.edu (C. Beeler).

Funding/support: This work was not publically funded but was provided financial support through the Infection Prevention Department at IU Health and IU Health System Decision Support.

Conflicts of interest: All authors have submitted the Disclosure of Potential Conflicts of Interest form to the institutional review board. L.T. and S.N. work for Health Catalyst, which provided machine learning, data engineering, and clinical subject matter support for model development. C.B., L.D., D.W., K.K., A.B., A.J-M., P.M., N.F., and J.A. have no conflicts of interest to disclose.

Author contributions: C.B. was the primary author responsible for drafting and finalizing the manuscript. L.D, K.K., J.A., and D.W. provided equal critical review and revision. A.B. provided independent statistical validation of the receiver operating characteristic curve. D.D., K.K., L.T., A.J.M., P.M., and N.F. helped design the model and provided final edits to the manuscript. All authors have seen and approved the manuscript and contributed significantly to the work. This manuscript has not been previously published nor is it being considered for publication elsewhere.

Central line-associated bloodstream infections (CLABSIs) are a major cause of healthcare-associated infections (HAIs) and contribute to increased morbidity, length of hospital stay, and cost.[1,2] The U.S. Centers for Disease Control and Prevention estimates that approximately 80,000 new CLABSIs occur in the United States every year, and data show a 12%-25% increased risk of mortality in hospitalized patients who develop a CLABSI.[3,4]

Advances have been made in understanding the pathogenesis of CLABSIs and implementing infection prevention bundles to reduce the incidence of infection, including attention to appropriate dressing, tubing changes, skin care, and hub disinfection. In the past decade, these efforts have resulted in a 60% reduction in rates of CLABSIs and $414 million in cost savings.[4] Other strategies that have been validated to reduce the rate of CLABSIs include using ethanol lock solutions, physician-targeted educational interventions, and

0196-6553/
https://doi.org/10.1016/j.ajic.2018.02.021

infection preventionist (IP) monitoring for central venous line (CVL) bundle adherence.[5-7]

Despite progress on a limited scale, the incidence of CLABSIs remains high and continues to contribute to increased patient morbidity and mortality, as well as millions of dollars in potentially avoidable healthcare costs.[8,9] Staff time, fluctuating patient volume, patient flow, and organizational priorities have been cited as barriers to providing infection prevention interventions in a consistent and timely manner.[8] In addition, despite a good understanding of the risk factors associated with the development of CLABSIs, healthcare organizations do not have a reliable way to identify in real time patients who are at higher risk for infection. An overstretched IP workforce is pressed to synthesize multiple data points in patient records to determine who is at highest risk and needs individualized time allocated to preventative efforts. If these high-risk patients could be identified in a systematic fashion, certain interventions, such as ethanol locking, intensive IP rounding, and physician notification for more urgent line removal, that are not otherwise feasible for the entire population with CVLs could be implemented to help prevent infection.

Machine learning has been recently popularized as a statistical technique to use relationships found in existing datasets to make predictions about future events. It uses multiple different techniques (such as boot strapping, cross-validation, generalized linear models, and decision trees or random forest) to develop a descriptive model of the data. This technique has already been used in healthcare settings to predict reservoirs for zoonotic diseases,[10] predict clinical outcomes from Ebola virus,[11] and predict mortality in patients with septic shock.[12] This strategy could be used in patients with a CVL to identify in real time those at highest risk by developing a risk score based on interactions between clinical variables found in the electronic medical record (EMR). Healthcare risk prediction models for CLABSIs based on a proportional hazards model developed by a statistician have been previously described.[13] Models developed by machine learning, however, are in their infancy, and their development has not been studied in depth. Other authors have explored the use of this technology predominantly as related to prediction of *Clostridium difficile* infection.[14-17] The main difference between models designed by a statistician and machine learning models is that the latter have no preconceived hypotheses about relationships between variables and outcomes. Machine learning is focused on optimizing predictive accuracy, whereas statistical techniques used by a statistician attempt to infer how much a variable contributes to the explanatory power of a model. Machine learning is also more concerned with the production of a model that can be implemented in the current work flow using the available data rather than testing hypotheses about relationships. Methods for model derivation within machine learning (such as random forest and logistic regression when used to optimize predictive accuracy) are further separated in that the former relies on a linear decision boundary in the data to separate the positive and negative outcomes (i.e., CLABSI present or CLABSI absent). The random forest, on the other hand, is an ensemble learning method that uses bootstrapped aggregation to leverage the results of many shallow decision trees, each based on a subset of variables and patient encounters. The fact that it is a tree-based algorithm allows the random forest to not only handle interactions between the model input features (i.e., the independent variables) in a way that logistic regression cannot, but also to discover nonlinear decision boundaries in the data. These traits allow the random forest to often better optimize the predictive power of the model. In this study, we describe the process of development and validation of a CLABSI risk prediction model at Indiana University Health Academic Health Center (IUH AHC). Considering that controlled trials around this new technology are scarce, this manuscript is focused on methods surrounding the creation, delivery, and accuracy of patient risk scores for CLABSIs.

METHODS

Design and setting

We performed a retrospective case-control analysis of adult, neonatal, and pediatric patients admitted to IUH AHC between January 1, 2013, and May 31, 2016, who had a CVL (internal jugular, subclavian, femoral, tunneled, non-tunneled, peripherally inserted, umbilical, or port). IUH AHC consists of 3 tertiary-care hospitals (2 adult and 1 pediatric) in Indianapolis, Indiana, with 1,238 licensed beds and 52,000 annual admissions.

Selection of cases and controls

Machine learning algorithms build models by learning patterns between attributes and outcomes in a particular dataset. Here, this was done by looking at the attributes of past patients who did and did not get a CLABSI as defined by the National Healthcare Safety Network (NHSN)[18] and excluding those who met Mucosal Barrier Injury criteria. Surveillance was performed by 1 IP who was certified in infection prevention and control. All data from patients with a CVL from January 1, 2013, to May 31, 2016, were used in the model generation.

Data collection

An interdisciplinary team developed the CLABSI risk prediction model. The team was composed of clinical experts including physicians, IPs, analysts, statisticians, health service researchers, a decision support team, and representatives from a healthcare analytics corporation (Health Catalyst). The team suggested evidence-based CLABSI risk factors that are thought to contribute to an increased risk of CLABSIs as well as some hypothesized risk factors and demographic information as a starting point. While machine learning processes excel at discovering which previously unappreciated variables are important in predicting a particular outcome, one cannot practically pull in the entire EMR. As such, we began the variable selection process by analyzing all variables with a reasonable backing in the literature and iterated from there based on feedback from a model validation process. These initial fields included:

- Intrinsic risk factors: age, sex, history of CLABSI, history of immunodeficiency in general, HIV, leukemia, lymphoma, and neutropenia (all by coded data).
- Extrinsic risk factors: chlorhexidine bathing non-adherence (defined as cumulative missed days), routine bathing non-adherence (defined as cumulative missed days), days in hospital prior to placement of the CVL, device days of having any CVL, device days for specific CVLs (peripheral, internal jugular, port, femoral, tunneled, or non-tunneled), device days cumulative of all CVLs, and parenteral nutrition.

Since the dataset is based on individual patient encounters, a given patient profile contains a day-count specific to each of the CVLs used thus far in the encounter as well as the cumulative count of all CVL days during the encounter. Parenteral nutrition is also based on a day-count for the entire encounter, even if the patient was not being fed intravenously on the day the risk prediction was made. Similarly, the patient's CVL-specific day-count persists for any CVL that has been removed. The patient receives a risk score if they have any CVL in place that day.

The data sources included nurse charting notes, physician diagnosis fields, DRG coded data, and ICD-10 encoded data. Data were