**SUPPLEMENTAL MATERIAL**

**Table S.1.** Carbapenemase Genes in Carbapenemase-Producing Admission Isolates

|  |  |  |
| --- | --- | --- |
| Ambler Class Carbapenemase and Overall Percentage (%) | Species | Number of Isolates with Molecularly-Confirmed Carbapenemases\*(n=23) |
| Class A – KPC (52%) | *Klebsiella pneumoniae* | 8 |
|  | *Enterobacter cloacae* | 2 |
|  | *Escherichia coli* | 1 |
|  | *Citrobacter amalonaticus* | 1 |
| Class B – NDM (9%) | *Klebsiella pneumoniae* | 1 |
|  | *Citrobacter amalonaticus* | 1 |
|  – VIM (4%) | *Pseudomonas**aeruginosa* | 1 |
| Class D – OXA-23 (9%) | *Acinetobacter baumannii* | 1 |
|  | *Acinetobacter radioresistans* | 1 |
|  – OXA-24 (9%) | *Acinetobacter baumannii* | 2 |
|  – OXA-48-like (4%) | *Klebsiella pneumoniae* | 1 |
| Classes B and D – NDM + OXA-48-like (13%) | *Klebsiella pneumoniae* | 3 |

**\*** 37 isolates were positive for carbapenemase production by the modified carbapenem inactivation method (mCIM) test and identified as carbapenemase-producing organisms (CPOs), of which 23 were positive for one or more carbapenemase genes on the CheckPoints assay. The disposition of the 14 CPO isolates without confirmed carbapenemases based upon Check-MDR CT103XL assay results included: *Enterobacter cloacae* (4; all AmpC-positive); *Acinetobacter baumannii* (4); *Citrobacter amalonaticus* (1); *Achromobacter xylosoxidans* (1; ESBL-positive); and four unevaluated isolates (*Aeromonas* spp. (2), *Enterobacter cloacae* (1); *Escherichia coli* (1)).

**CART Decision Trees for Predicting CRE Colonization at Hospital Unit Admission:**

We fit separate CART decision trees to predict CRE colonization because due to differences in underlying organism distribution and/or resources, institutions may wish to target CRE, rather than all CRO, colonization. Moreover, because perirectal swabs may be less sensitive for detecting NFCRO colonization, restricting to the outcome of CRE colonization could reduce outcome misclassification and improve model performance.

The decision tree to predict CRE colonization performed similarly, albeit slightly inferior to, the tree for predicting CRO colonization. It included a single variable, prior history of a CRE-positive culture in the six months preceding unit admission. Model performance improved significantly when the tree was tuned to increase sensitivity, but with a large increase in model complexity (12 included variables) and evidence of overfitting (e.g., a 16 percentage-point reduction in discrimination in random forests analysis compared to the single CART decision tree), suggesting that the tree would not generalize well to external data (Table S.2). Restricting to the outcome of CRE colonization did not improve model performance relative to prediction models for CRO and CPO colonization.

**Table S.2.** Performance Metrics of CART Decision Trees for Predicting CRE Colonization at Unit Admission

|  |  |  |
| --- | --- | --- |
|  | Raw (No Tuning) | Tuned to Increase Sensitivity |
| No. of Included Variables | 1 | 12 |
| Sensitivity | 5.8% | 70.2% |
| Specificity | 99.9% | 72.3% |
| Positive Predictive Value (PPV) | 64% | 10% |
| Negative Predictive Value (NPV) | 96% | 98% |
| C-Statistic | 0.53 | 0.76 |
| C-Statistic in Random Forests (RF) Analysis | 0.60 | 0.60 |

**Sensitivity Analyses:**

1. **Random Forests Analyses**

Random forests (RF) analysis is similar to CART analysis, or other tree-fitting algorithms, except that it generates many bootstrapped decision trees [1, 2]. Its output is less easily interpretable because it does not produce a singular, final tree, but as an ensemble method it generally improves accuracy and reduces model overfitting, i.e., increases generalizability. It also estimates the most important variables for predicting a given outcome [3, 4] (Figure S.1). We performed random forests analyses to fit 1,000 bootstrapped classification trees, using the randomForest package (version 4.6-14).

The C-statistic for predicting CRO colonization in RF analysis was 0.65, a 14% increase from the single decision tree (C-statistic 0.57). Consistent with the single decision tree’s placement of a recent CRO-positive culture in the root node, which is reserved for the most discriminatory variable, RF analysis also identified this exposure as the most important for predicting CRO colonization status at admission (Figure S.1).

Using the RF variable rankings, we selected the five most important, non-nested variables for inclusion in a multivariable logistic regression model (nested variables, e.g., prior CRO culture and prior MDRGN culture, would violate logistic regression’s requirement for variable non-collinearity). These variables in order of importance were: CRO-positive culture in the prior six months, total days of hospitalization in the prior six months, Elixhauser severity of illness score, and total DDD-standardized doses of antibiotics with Gram-negative coverage and immunosuppressive therapy in the prior three months. The resulting model’s C-statistic was 0.62 (Figures S.1 and S.2). Although we did not pursue this analysis further, this logistic regression model could be evaluated for conversion to a more user-friendly prediction model (e.g., a clinical risk score).

In the RF model for CPO colonization, discrimination also rose (C-statistic 0.70, 21% improvement) relative to the original CART decision tree. The top five, non-nested variables in order of importance for predicting CPO colonization were: CRO-positive culture in the prior six months, Elixhauser severity of illness score, total DDD-standardized doses of antibiotics with Gram-negative activity in the prior three months, time from hospital-to-unit admission (days), and total number of days of immunosuppressive therapy in the three months preceding unit admission. In logistic regression incorporating these variables, the C-statistic was 0.70 (Figures S.2 and S.3).

**Figure S.1.** Random Forests Analysis Variable Importance Plot of Most Predictive Variables for CRO Colonization at Unit Admission and Corresponding Logistic Regression Output

****

**A).** Random forests-generated variable importance plot ranking the ten most predictive variables for CRO colonization; **B).** Output from multivariable logistic regression analysis incorporating the top five, non-collinear variables from (A).

**Figure S.2.** Receiver Operating Characteristic (ROC) Curves for Multivariable Logistic Regression Models Predicting CRO and CPO Colonization at Unit Admission, Incorporating the Five Most Predictive Variables Identified in Random Forests Analysis.

****

**A).** Receiver operating characteristic (ROC) curve for a multivariable logistic regression model with the outcome of carbapenem-resistant organism (CRO) colonization at unit admission and five independent variables: CRO-positive culture in the prior six months, total days of hospitalization in the prior six months, Elixhauser severity of illness score, total defined daily dose (DDD)-standardized doses of antibiotics with Gram-negative coverage in the prior three months, and total DDD-standardized doses of immunosuppressive therapy in the prior three months; **B).** Corresponding ROC curve for the outcome of CPO colonization and five independent variables: CRO-positive culture in the prior six months, Elixhauser severity of illness score, total DDD-standardized doses of antibiotics with Gram-negative activity in the prior three months, time from hospital-to-unit admission (days), and total number of days of immunosuppressive therapy in the prior three months.

**Figure S.3.** Random Forests Analysis Variable Importance Plot of Most Predictive Variables for CPO Colonization and Corresponding Logistic Regression Output

****

**A).** Random forests-generated variable importance plot ranking the ten most predictive variables for CPO colonization; **B).** Output from multivariable logistic regression analysis incorporating the top five, non-collinear variables from (A).

1. **Re-Fit CART Trees to Up-Weight Sensitivity**

Due to the large imbalances (i.e., rare outcomes) in our data, we refit CART trees to increase sensitivity by imposing a greater “cost” for misclassifying colonized patients as negative by incorporating CRO and CPO outcome prior probabilities of 0.50 and 0.10, respectively. Overall discrimination improved in resulting models, as did sensitivity (sensitivities increased to 62% and 47%, respectively), but with attendant increases in model complexity and likely decreases in practical utility (Figures S.4 and S.5).

**Figure S.4.** Decision tree for predicting CRO perirectal colonization at hospital unit admission, fit with 50% priors to up-weight tree sensitivity**.** Gray-shaded terminal nodes indicate that the tree would classify patients as CRO-colonized, and accompanying percentages reflect the probability that patients assigned to a given terminal node are CRO-positive. Terminal node numbering is included in parentheses. The tree possessed a sensitivity of 61.8%, a specificity of 70.4%, a positive predictive value of 14.5%, and a negative predictive value of 95.8%. Its C-statistic was 0.70, and its C-statistic in random forests analysis was 0.65.

****

**Figure S.5.** Decision tree for predicting CPO perirectal colonization at hospital unit admission, fit with 10% priors to up-weight tree sensitivity.Gray-shaded terminal nodes indicate that the tree would classify patients as CPO-colonized, and accompanying percentages reflect the probability that patients assigned to a given terminal node are CPO-positive. Terminal node numbering is included in parentheses. The tree possessed a sensitivity of 47.2%, a specificity of 99.1%, a positive predictive value of 40.5%, and a negative predictive value of 99.3%. Its C-statistic was 0.75, and its C-statistic in random forests analysis was 0.69.

****

1. **Refit CART trees restricting to first, unique-patient encounters (n=2165)**

Due to multiple ICU admissions during the study period, our cohort included repeat-patient encounters. Time-fixed variables (e.g., race) could potentially be over-represented in models if patients were admitted many times during the study period. Because traditional CART approaches do not offer simple approaches like logistic regression (e.g., general estimating equations) to account for multiple observations from the same subject, we refit trees restricting to first, unique-patient encounters in order to evaluate changes in model performance.

The tree for predicting CRO colonization performed similarly when restricting to first patient encounters, but the tree for predicting CPO colonization failed to branch, possibly due to insufficient sample size in this restricted subset (only 20 CPO outcomes) (Table S.3).

**Table S.3.** Performance Metrics of CART Decision Trees for Predicting CRO and CPO Colonization at Unit Admission, Restricted to First Unique-Patient Admission Encounters During the Study Period.\*

|  |  |  |
| --- | --- | --- |
|  | CRO Decision Tree | CPO Decision Tree |
| No. of Outcomes (N=2165) | 141 | 20 |
| No. of Included Variables | 3 | 0 (Failed to Branch) |
| Sensitivity | 8.5% | NA |
| Specificity | 99.8% | NA |
| Positive Predictive Value (PPV) | 75% | NA |
| Negative Predictive Value (NPV) | 94% | NA |
| C-Statistic | 0.55 | 0.50 |

**\*** Number of first, unique-patient encounters during the study period equals 2165. Abbreviations: “NA”, not applicable.

1. **Re-performed both CART and RF analyses after restricting the CPO outcome to only isolates with molecularly-confirmed carbapenemases.**

In order to address possible outcome misclassification, for the outcome of CPO colonization we refit a decision tree and performed random forests analysis restricting to CPO isolates with molecularly-confirmed carbapenemases. Of the 36 CPO-positive swabs, 22 swabs were positive for one or more carbapenemases on the Checkpoints assay (representing 23 isolates, due to CP-CRE co-colonization on one swab). A decision tree fit to this data failed to branch, indicating that no variables were sufficiently predictive for this outcome. We also refit a decision tree with 10% priors in order to increase tree sensitivity. This tree possessed a C-statistic of 0.94 and included 14 variables, evidencing that it was overfit. In random forests analysis (no adjustment of priors), the C-statistic was 0.68, similar to the value for the primary CPO outcome (C-statistic 0.70).

Restricting the CPO outcome to swabs possessing molecularly-confirmed carbapenemases did not materially improve, or in some cases reduced, model performance.

**Patient Data**

Patient data was based on the day of, or days/weeks prior to, unit admission, and included the following: (a) demographic data (e.g., age, sex, race); (b) hospitalization encounter-level data (e.g., admission type, pre-admission location); (c) pre-existing medical conditions and co-morbidities; (d) multidrug-resistant organism (MDRO) colonization or infection (≤ 6 mos.), and contact precautions and VRE colonization status at unit admission; (e) days of inpatient and/or ICU hospitalization (≤ 6 mos.); (f) discharge to a post-acute care facility (long-term acute care hospital, skilled nursing or rehabilitation facility) (≤ 6 mos.); (g) indwelling hardware (≤ 3 mos.); (h) days and total defined daily dose (DDD)-adjusted doses [5] of immunosuppressive therapy, proton-pump inhibitors (PPIs), or histamine H2-receptor antagonists (H2-blockers) (≤ 3 mos.); (i) days and total DDD-adjusted doses of Gram-negative active antibiotic therapy (≤ 3 mos.); (j) invasive abdominal procedures or surgeries (≤ 3 mos.); and (k) recent international exposure, including hospitalization in another country in the prior six months and/or foreign travel of the patient or a partner in the prior 21 days, both assessed by standard nursing intake questionnaire. Consistent with CDC guidance [6], Johns Hopkins Healthcare System policy refers patients with recent international hospitalization (<6 mos.) for CRE peri-rectal colonization surveillance screening. We also calculated a 30-category, unweighted Elixhauser Comorbidity Index score (including both primary diagnoses and comorbidities) from the EMR [7].

**Supplemental Material References**

1. Chen X, Ishwaran H: Pathway hunting by random survival forests. *Bioinformatics* 2013, 29(1):99-105.

2. Strobl C, Malley J, Tutz G: An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 2009, 14(4):323-348.

3. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT: Application of random forests methods to diabetic retinopathy classification analyses. *PloS one* 2014, 9(6):e98587.

4. Rose S: Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol* 2013, 177(5):443-452.

5. Defined Daily Dose (DDD) [<http://www.who.int/medicines/regulation/medicines-safety/toolkit_ddd/en/>]

6. CDC: Facility Guidance for Control of Carbapenem-resistant Enterobacteriaceae (CRE) – November 2015 Update CRE Toolkit. In*.* Atlanta, Georgia: CDC; 2015.

7. Elixhauser A, Steiner C, Harris DR, Coffey RM: Comorbidity measures for use with administrative data. *Medical care* 1998, 36(1):8-27.