Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients*

Ryan J. Delahanty, PhD[1]; David Kaufman, MD, FCCM[2]; Spencer S. Jones, PhD[1]

Objectives: Risk adjustment algorithms for ICU mortality are necessary for measuring and improving ICU performance. Existing risk adjustment algorithms are not widely adopted. Key barriers to adoption include licensing and implementation costs as well as labor costs associated with human-intensive data collection. Widespread adoption of electronic health records makes automated risk adjustment feasible. Using modern machine learning methods and open source tools, we developed and evaluated a retrospective risk adjustment algorithm for in-hospital mortality among ICU patients. The Risk of Inpatient Death score can be fully automated and is reliant upon data elements that are generated in the course of usual hospital processes.

Setting: One hundred thirty-one ICUs in 53 hospitals operated by Tenet Healthcare.

Patients: A cohort of 237,173 ICU patients discharged between January 2014 and December 2016.

Design: The data were randomly split into training (36 hospitals), and validation (17 hospitals) data sets. Feature selection and model training were carried out using the training set while the discrimination, calibration, and accuracy of the model were assessed in the validation data set.

Measurements and Main Results: Model discrimination was evaluated based on the area under receiver operating characteristic curve; accuracy and calibration were assessed via adjusted Brier scores and visual analysis of calibration curves. Seventeen features, including a mix of clinical and administrative data elements, were retained in the final model. The Risk of Inpatient Death score demonstrated excellent discrimination (area under receiver operating characteristic curve = 0.94) and calibration (adjusted Brier score = 52.8%) in the validation dataset; these results compare favorably to the published performance statistics for the most commonly used mortality risk adjustment algorithms.

Conclusions: Low adoption of ICU mortality risk adjustment algorithms impedes progress toward increasing the value of the healthcare delivered in ICUs. The Risk of Inpatient Death score has many attractive attributes that address the key barriers to adoption of ICU risk adjustment algorithms and performs comparably to existing human-intensive algorithms. Automated risk adjustment algorithms have the potential to obviate known barriers to adoption such as cost-prohibitive licensing fees and significant direct labor costs. Further evaluation is needed to ensure that the level of performance observed in this study could be achieved at independent sites. (*Crit Care Med* 2018; 46:e481–e488)

Key Words: critical care; ICU scoring systems; machine learning; mortality risk

Reliable risk adjustment algorithms are essential for evaluating the performance of healthcare providers (1, 2). Such algorithms facilitate comparisons between facilities, allow for the establishment of performance benchmarks, and enable healthcare providers to set goals for quality improvement (3, 4). A number of risk adjustment algorithms have been specifically developed to estimate and adjust for ICU patients' risk of mortality (1, 5–7). Some of these algorithms are in their third or fourth iteration and have been improved over the course of decades. The most documented mortality risk adjustment algorithms include the Acute Physiology and Chronic Health Evaluation (APACHE), the Simplified Acute Physiology Score (SAPS), and the Mortality Probability Model (MPM). The discrimination of each of these models, as measured by the area under a receiver operating characteristic curve

*See also p. 1024.

[1]Tenet Healthcare, Nashville, TN.

[2]The Intensivist Group/Sound Physicians, Tacoma, WA.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Dr. Kaufman received funding from Alesky Belcher LLC, and he has consulting agreements with Intensix Inc, Natanya, Israel, and Advanced ICU, St. Louis, MO (no financial payments received from either in the past 36 mo). The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: spencer.jones@tenethealth.com

DOI: 10.1097/CCM.0000000000003011

(AUC), ranges between 0.81 for MPM and 0.89 for APACHE (1, 2, 8). Although there is no "gold standard," APACHE consistently demonstrates the highest performance in terms of AUC (4, 9).

Even though there are a number of validated and accepted risk adjustment algorithms available, and there is a compelling value proposition for their use, utilization of these risk adjustment algorithms is low (10). Recent estimates suggest that only 12% of ICUs use some type of mortality risk adjustment algorithm (3). Low levels of adoption may be attributable to two primary factors. First, although the underlying algorithms for APACHE-IV and MPM_0-III are in the public domain and available at no cost, to use these tools in practice typically requires the payment of licensing, implementation, and maintenance fees which may be cost prohibitive for many hospitals. Second, implementations of these algorithms often require critical care clinicians to engage in time-intensive collection and documentation of patient data that is not captured in typical critical care workflows nor is readily available in commonly used clinical information systems (4, 8).

Hospital adoption of electronic health record (EHR) systems has accelerated in response to federal incentive programs, and nearly 97% of hospitals now have at least a basic EHR (11). The widespread adoption of EHRs containing troves of physiologic measurements, information about treatment, patient demographics, diagnostics, and medical history has the potential to obviate the need for risk adjustment algorithms that require manual or custom data collection. Both the Veterans Health Administration and Kaiser-Permanente have demonstrated that fully automated algorithms can perform comparably to risk adjustment algorithms that require specialized manual inputs (12, 13).

In this article, we describe the development and evaluation of a risk adjustment algorithm using modern machine learning methods that rely on data that are routinely and passively collected in the course of patient care and hospital operations. The Risk of Inpatient Death (RIPD) score uses data elements that can be extracted from clinical and administrative systems in an automated fashion and therefore requires no manual chart review, abstraction, or custom data entry. This algorithm, like other commonly used mortality risk adjustment tools, is designed to be used for retrospective risk adjustment, rather than prospective prediction of patients' risk of mortality. We compare the performance statistics generated in this analysis to performance statistics reported in the literature for APACHE, SAPS, and MPM.

MATERIALS AND METHODS

Hospital and ICU Selection

This study was deemed to meet the conditions for institutional review board (IRB) exemption by an external IRB (WIRB, Puyallup, WA). At the time of this analysis, Tenet Healthcare operated 79 acute care hospitals across the United States. From this group, hospitals were chosen based on four criteria: 1) had one or more defined ICUs meeting adult critical care criteria according to National Healthcare Safety Network guidelines as published by the Centers for Disease Control and Prevention (14), 2) were not children's hospitals, 3) saw a minimum of 100 ICU patients in the study period, and 4) used the Cerner Millennium EHR system. In total, 53 hospitals and 131 ICUs met these criteria.

Patient Selection

All inpatients 18 years old or older discharged between the beginning of January 2014 and December 2016 that spent a portion of their hospital encounter in an ICU were included in this analysis. Patients with unknown discharge disposition, or orders indicating they were to receive "comfort measures only" prior to entry to the ICU, or those seen in a pediatric setting at any point during their encounter, were excluded from the analysis (n = 425; 0.2%). In total, there were 237,173 patients included in the analysis. Patient characteristics are described in **Table 1**.

Data Sources

Data came from two primary sources: 1) clinical data including laboratory test results, vital signs, and documentation of notable events or procedures (e.g., mechanical ventilation) and 2) administrative data including demographic, billing diagnoses, and utilization history. Clinical data were captured as part of the usual processes of care in one of six instantiations of the Cerner Millennium EHR system and archived in a common enterprise data warehouse. Administrative data were also captured through the usual course of hospital operations and archived in the same data warehouse.

Feature Selection

There are more than 15,000 different clinical variables available in our EHR's results database alone; in addition, administrative data include thousands of features that could potentially be incorporated into a risk adjustment algorithm. Given the breadth of options, there was an immediate need to reduce this feature space to a more manageable size that would allow us to produce a risk adjustment algorithm that was feasible from both a computational and implementation standpoint. This feature selection process was carried out in three stages: first, an environmental scan of existing risk adjustment algorithms and of published ICU admission criteria (15); second, consultation with one of the authors (D.K.) who is a board-certified critical care physician; and third, automated feature selection via machine learning algorithms, a process that is further described below.

Administrative data related to patient demographics, utilization history, and diagnoses were considered as potential features. Demographic features included patient age, sex, race, language, admission through the emergency department, and indication of surgical admission. Diagnosis-related features included All Patient Refined-Diagnosis Related Group (APR-DRG) codes as well as Medicare cost weights for patient Medicare Severity-Diagnosis Related Groups (MS-DRGs) (16–18).

Clinical observations were limited to those proximate to the time of admission to the ICU with a minimum of 24 hours before admission and a maximum of 24 hours postadmission. In some cases, we used first or last measurements (e.g., last Glasgow Coma Scale Score), changes in measurements or measurement

TABLE 1. Descriptive Statistics for the Training and Validation Data Sets[a]

Variable	Training Set (*n* = 82 ICUs)	Validation Set (*n* = 49 ICUs)	Overall (*n* = 131 ICUs)
Patients, *n*	146,982	90,191	237,173
Age (yr), mean (SD)	64.3 (17.8)	62.9 (17.6)	63.7 (17.8)
Female sex, *n* (%)	68,171 (46.4)	41,571 (46.1)	109,742 (46.3)
Race, *n* (%)			
Black	27,171 (18.5)	18,292 (20.3)	45,463 (19.2)
White	75,963 (51.7)	40,748 (45.2)	116,711 (49.2)
Other/unknown	43,858 (29.8)	31,151 (34.5)	74,999 (31.6)
Language English, *n* (%)	129,407 (88.0)	74,803 (82.9)	204,210 (86.1)
Deaths, *n* (%)	13,725 (9.3)	8,168 (9.1)	21,893 (9.2)
Length of stay, mean days (SD)	7.6 (9.8)	8.0 (11.2)	7.8 (10.3)
Time until ICU admission, mean days (SD)	0.8 (2.7)	0.8 (2.8)	0.8 (2.7)
30-d readmission return patients, *n* (%)	19,897 (13.5)	10,916 (12.1)	30,813 (13.0)
Emergency department admit, *n* (%)	121,607 (82.7)	71,716 (79.5)	193,323 (81.5)
Transfer patients, *n* (%)	15,950 (10.9)	11,732 (13.0)	27,682 (11.7)
Surgical inpatients, *n* (%)	28,029 (19.1)	20,720 (23.0)	48,749 (20.6)
Selected All Patient Refined-Diagnosis Related Groups, *n* (%)			
720—septicemia	13,531 (9.2)	8,178 (9.1)	21,709 (9.2)
133—pulmonary edema and respiratory failure	5,614 (3.8)	2,324 (2.6)	7,938 (3.3)
710—infectious and parasitic diseases with operating room procedure	2,849 (1.9)	1,783 (2.0)	4,632 (2.0)
130—respiratory diagnosis with vent 96+ hr	1,926 (1.3)	897 (1.0)	2,823 (1.2)
044—intracranial hemorrhage	2,030 (1.4)	1,635 (1.8)	3,665 (1.5)
ICU type, *n* (%)			
General ICU	54 (65.9)	30 (61.2)	84 (64.1)
Cardiovascular ICU	12 (14.6)	10 (20.4)	22 (16.8)
Surgical ICU	10 (12.2)	8 (16.3)	18 (13.7)
Trauma ICU	6 (7.3)	1 (2.0)	7 (5.3)

[a]Feature selection and model training was conducted using the training set, while model performance was evaluated using the validation set.

times (e.g., heart rate), or measurement means (e.g., respiratory rate). In addition, we used literature to guide feature engineering of some novel feature interactions, for example, shock index (heart rate/systolic blood pressure) by age (19).

Stages 1 and 2 of the feature selection process yielded a significantly reduced feature set that included only 215 clinical and administrative features. These features were carried forward to stage 3, supervised selection process described below.

Model Development and Evaluation

The objective of the model was to reliably estimate ICU patients' risk of in-hospital death. To this end, XGBoost (v. 0.4; T. Chen, T. He, M. Benesty; https://CRAN.R-project.org/package=xgboost), a freely available, open-source software library for machine learning was used to build a model that could reliably estimate the risk of inpatient mortality (20, 21). The model was trained using patients from 36 of the 53 hospitals (146,982 patients). Patients from the remaining 17 hospitals (90,191 patients) were "held out" to serve as a validation set. Assignment to the training versus validation set was determined via simulation study in which we generated 200 random splits of the hospital population. We then calculated observed/expected mortality rates for the training and validation sets for each random split; we then selected the split that minimized the difference in observed/expected mortality between the training and validation sets. The objective of this simulation exercise was to reduce the risk of bias that could occur if a disproportionate share of low or high mortality ICUs were assigned to the training or validation set.

Model development was carried out using the XGBoost package within the R Software for Statistical Computing (R Foundation for Statistical Computing, c/o Institute for Statistics and Mathematics, Vienna, Austria; http://www.r-project.org). R is a free and open-source programming language and environment for statistical computing and graphics. XGBoost models are comprised of thousands of relatively simple decision trees. These models are trained iteratively by combining individual decision tree models to optimize on an evaluation metric, such as AUC. At each iteration, an additional decision tree is added to the "ensemble" of previously trained decision trees. However, each new decision tree takes into account errors made in the previous iterations. In this way the model "learns" its own shortcomings and introduces a new decision tree to address those shortcomings. One of the benefits of these models over regression-based methods is that they are robust in the presence of missing data and thus do not require imputation. As such, we provided extreme negative values to indicate missingness (−9999), which in this case yielded results superior to median value imputation. We used a step size shrinkage of 0.3, which controls the rate of optimization in the model. We allowed decision tree complexity to reach a maximum depth of three. To avoid overfitting, we used five-fold cross validation in the training dataset and drew subsamples of 80% of patient encounters in training.

Here we could criticize the way they selected the features !

In total, 215 features were entered into the supervised machine learning selection process. Relative influence of the model features was determined by evaluating the "gain," a value based on the number of times a feature is selected for splitting across the entire ensemble of decision trees. Gain is the increase in accuracy brought about by including a feature to the branches of the decision tree. The relative influence of a feature is averaged over all trees. Initially, all features were entered in to the model, several rounds of feature selection ensued. At each round, the bottom third of the feature set (in terms of gain) was dropped from the model. This process was carried out until we observed substantial declines in model discrimination (in terms of AUC). Ultimately 17 features were retained in the final model.

Could be criticized!

Model performance was assessed based on discrimination, accuracy, and calibration. Model discrimination was assessed via the AUC. The Hosmer-Lemeshow test is often used to assess the calibration of risk; however, this test can be unreliable when the sample size is large (22, 23). Instead we assess accuracy using the adjusted Brier score and visual analysis of calibration curves with observed and expected deaths plotted across each risk decile. Standardized mortality ratios (SMRs) were calculated for the training and validation sets as the number of observed deaths divided by expected deaths (sum of RIPD scores for all patients in a given ICU). The average SMR and 95% CI were calculated using standard methods. Specifically, we used the summary SE function with the Rmisc package within the R statistical software.

RESULTS

Hospital and Patient Characteristics

The 53 hospitals included in our study were diverse, in that they are located in 14 different states and differ significantly in size (minimum bed count = 41, maximum bed count =762, median bed count = 217, interquartile range = 153–400). The majority of these hospitals are located in metropolitan areas; however, three hospitals are located in micropolitan areas, and one hospital is located in a rural area. Likewise, the majority (47) of these hospitals are nonteaching community hospitals. There were 131 individual ICUs across the 53 hospitals, 84 were general ICUs, 22 were cardiovascular ICUs, 18 were surgical ICUs, and seven were trauma ICUs.

Descriptive statistics for the study population are shown in Table 1. The mortality rate observed across all hospitals for the study period was 9.2%. Patients' age, sex, race, and language varied widely across the study hospitals. Additionally, there was substantial variability in admission source and length of stay across the study hospitals. Notably, in-bound transfers from other healthcare facilities, often excluded from risk adjustment models, accounted for 11.7% of our study population.

RIPD Score Features

Feature selection was done by removing variables which contributed the least gain to the model. This was used to reduce the number of features from 215 to 17. The complete list of features retained in the model ranked by relative influence is described in **Table 2** (for a list of top 100 features and feature stability, see **Appendix Tables 1** and **2**, Supplemental Digital Content 1, http://links.lww.com/CCM/D222). The retained features included a mix of administrative data elements, laboratory test results, vital signs, and documentation of use of mechanical ventilation or supplemental oxygen. The most important features were based on commonly available administrative data (e.g., APR-DRG risk of mortality and severity of illness weights). Engineered variables for shock index and shock index by age (heart rate/systolic blood pressure × age) were important in the model. Physiologic variables including heart rate, pulse oximetry, and mean respiratory rate also ranked in the top 10 most influential features retained in the model. Somewhat intuitively, last values and mean values (for up to 24 hr after ICU admission) tended to provide better prediction than first values. Last evidence of using mechanical ventilation or any type of oxygen therapy was retained and coded as yes/no.

RIPD Score Calibration and Discrimination

To evaluate agreement between observed and expected mortality across risk strata, we used the adjusted Brier score. The adjusted brier score represents the percent reduction in deviation when using the RIPD model as opposed to assigning all patients a risk score equal to the average mortality for the entire population. A higher percentage reduction indicates better model accuracy (9). The adjusted Brier score for the RIPD score was 52.8%, which compares favorably to the scores reported for other risk adjustment models. The calibration curve showed good correspondence between predicted and observed risk deciles (**Fig. 1**). In terms of discrimination, the RIPD score AUC in the validation set was 0.94 with strong performance across all subgroups. The SMR for the train set was 1.01 (95% CI, 1.00–1.03) and 1.01 (95% CI, 0.99–1.03) for the test set (**Table 3**).

Important Results !

TABLE 2. Summary of Features Used in Risk of Inpatient Death Score

Feature	Mean (SD)	Missingness (%)	Relative Influence (%)
APR-DRG risk of mortality (integer 1–4)	2.5 (1.1)	6.19	24.76
Last Glasgow Coma Score (integer 1–15)	13.4 (3)	8.98	16.23
APR-DRG severity of illness (integer 1–4)	2.6 (1.1)	1.06	10.42
Medicare cost weight index	2.4 (2.4)	0.00	7.20
Last measured shock index[a]	43.7 (21.1)	0.26	6.14
Last shock index	0.7 (0.3)	0.26	5.87
Last pulse oximetry	96.1 (7.4)	0.44	4.03
Mean pulse oximetry	97 (2.6)	0.44	3.63
Last systolic blood pressure	123.5 (23.9)	0.15	3.47
Last heart rate	82.4 (20.1)	0.08	2.95
Mean respiratory rate	19.2 (3.8)	0.10	2.82
Last CO_2 measurement	24.5 (4.7)	3.26	2.66
Mean temperature (°F)	98.3 (0.8)	3.75	2.52
Last blood urea nitrogen	25.1 (20.4)	3.54	2.49
Change in creatinine level	–0.14 (0.83)	3.59	1.93
Last evidence of any oxygen therapy (yes/no)	59.2% (yes)	3.96	1.58
Last mechanical ventilation status (yes/no)	14.4% (yes)	3.96	1.29

APR-DRG = All Patient Refined-Diagnosis Related Group.
[a]Shock index = (heart rate/systolic blood pressure) × age.

DISCUSSION

The provision of healthcare in American ICUs costs over $80 billion per year, or 4% of all healthcare spending, and nearly 39% of all hospital costs in the United States (24, 25). The introduction and diffusion of new technologies continues to push these costs higher (26). There is a growing body of evidence that suggests that as critical care costs increase, much of this additional spending is wasteful and does not yield better outcomes for patients (27, 28). However, given the variation in patient populations across different ICUs, it is difficult to identify and reduce this waste and inefficiency (3). Accurate, generalizable, and easily implemented risk adjustment methods are necessary to enable healthcare providers and policy makers to measure and implement strategies to increase the value of the healthcare delivered in ICUs. Current levels of adoption of ICU mortality risk adjustment tools are insufficient to enable the large scale change that is needed. The RIPD score has many attractive attributes that address the key barriers to adoption of ICU mortality risk adjustment tools.

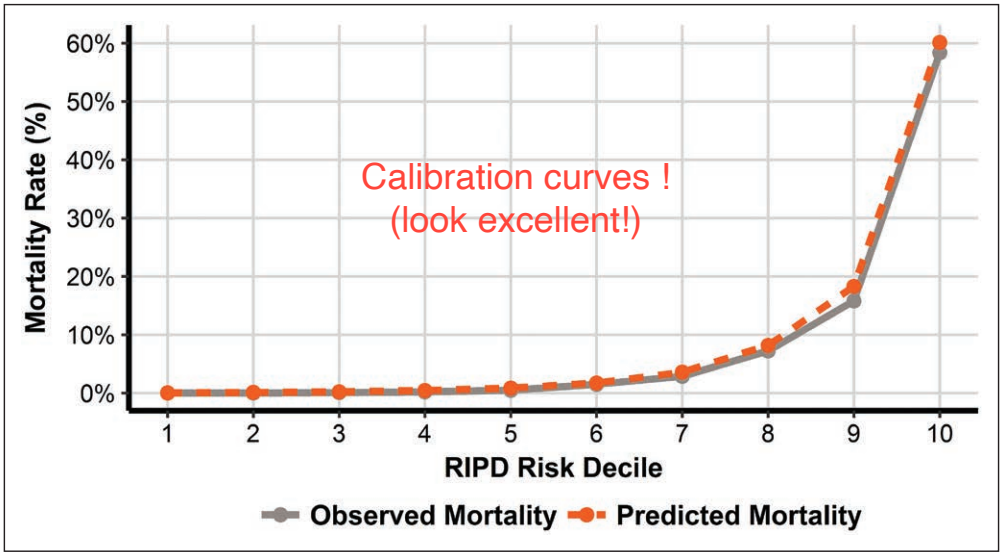

Figure 1. Calibration curves show observed and predicted mortality rates across deciles of the Risk of Inpatient Death (RIPD) score distribution. This figure shows close agreement between observed and predicted mortality rates across all deciles of the RIPD score.

First, we have demonstrated that the RIPD score is generalizable outside of hospitals where it was developed. As noted above, the RIPD algorithm was developed using data from 36 of the eligible 53 hospitals. This design allowed us to test the calibration and discrimination of the RIPD score in a set of hospitals

TABLE 3. Crude and Standardized Mortality Ratios, Accuracy, and Discrimination in Training Set, Testing Set, and Testing Set Subgroups

Dataset	Sample size	Deaths	Mortality Rate (%)	Standardized Mortality Ratio (95% CI)	Adjusted Brier Score (%)	Area Under a Receiver Operating Characteristic Curve
Training set (82 ICUs)	146,982	13,725	9.3	1.01 (1.00–1.03)	55.3	0.951
Validation Set (49 ICUs)	90,191	8,168	9.1	1.01 (0.99–1.03)	52.8	0.943
Top 5 All Patient Refined-Diagnosis Related Groups (validation set only)						
720—septicemia	8,178 (9.1)	2,013	24.6	0.98 (0.93–1.02)	59.0	0.892
133—pulmonary edema and respiratory failure	2,324 (2.6)	410	17.6	1.06 (0.97–1.17)	50.0	0.923
710—infectious and parasitic diseases with OR procedure	1,783 (2.0)	302	16.9	0.96 (0.85–1.07)	41.7	0.829
130—respiratory system diagnosis with ventilatory support 96+ hr	897 (1.0)	266	29.7	1.13 (1.00–1.27)	32.4	0.747
044—intracranial hemorrhage	1,635 (1.8)	338	20.7	1.01 (0.98–1.21)	72.4	0.973
ICU type (validation set only)						
General ICU	53,176 (59.0)	4,733	8.9	1.01 (0.98–1.04)	54.0	0.943
Cardiovascular ICU	18,438 (20.3)	732	4.0	0.97 (0.92–1.02)	55.3	0.941
Surgical ICU	16,769 (18.6)	531	3.1	1.04 (0.99–1.09)	46.9	0.942
Trauma ICU	1,808 (2.0)	154	8.5	1.00 (0.84–1.16)	46.7	0.958
Patient types (validation set only)						
Emergency department admit	71,716 (79.5)	6,824	9.5	1.00 (0.78–1.02)	52.2	0.941
Surgical patients	20,720 (23.0)	1,234	6.0	0.91 (0.86–0.96)	39.2	0.909
Transfer patients	11,732 (13.0)	1,602	13.7	1.05 (1.00–1.10)	50.1	0.906
30-d readmission patients	10,916 (12.1)	1,593	14.6	1.16 (1.11–1.22)	48.7	0.917

distinct from the development set. This design increases the likelihood that the risk adjustment algorithm is sufficiently robust and will work well for other hospitals.

Second, RIPD addresses cost, which is likely the most significant barrier to adoption of these tools (2). These costs come in the form of licensing, implementation, training, and most significantly, manual data collection (direct labor). An automated algorithm based on freely available, open-source software such as the RIPD score would effectively eliminate additional direct labor or licensing costs, which comprise the lion's share of the costs of maintaining an ICU mortality risk-adjustment system.

Third, the RIPD score does not require any custom data entry and relies entirely on data elements that are likely to be accessible to hospitals that have electronic clinical and administrative systems. Given the widespread adoption of EHRs, the RIPD score could be broadly and efficiently adopted for risk adjustment purposes.

Finally, healthcare providers may have been hesitant to adopt automated risk adjustment algorithms because transitioning from human-intensive algorithms to more automated risk adjustment algorithms has meant sacrificing performance (2, 8). For example, the more intensive APACHE-IV which requires, on average, 37 minutes of human effort per patient to calculate, has a reported AUC of 0.88, and outperforms the less human-intensive MPM_0-III (AUC = 0.82), which takes only about 11 minutes of human effort (8). The fully electronic SAPS (AUC = 0.82), an automated adaptation of the SAPS risk adjustment model, performs comparably to MPM but well below APACHE-IV (12). By contrast, the RIPD score (AUC = 0.94) is fully automated but does not kill any discriminatory performance, and in fact compares favorably to published statistics for APACHE-IV (4, 9).

LIMITATIONS

As noted earlier, administrative data elements, such as APR-DRG codes and their associated risk of mortality and severity scores, are important components of the RIPD score. The accuracy of administrative coding is dependent on the expertise of the coder; however, rigorous credentialing and coding regulations limit variation in the practice and accuracy of clinical

coding, and there is strong precedent for using APRDRG codes for risk adjustment (17, 29). We believe that the majority of hospitals use APR-DRG codes; however, hospitals that do not use APR-DRG codes will not be able to use the RIPD algorithm. To address this limitation, we have also developed and validated a version of the RIPD algorithm that does not incorporate APR-DRG or MS-DRG codes (RIPD_reduced). The RIPD_reduced algorithm also provides very good discrimination (AUC = 0.91) and calibration (adjusted Brier Score = 0.49). Further details related to the RIPD_reduced algorithm can be found in the **Appendix Tables 3** and **4** (Supplemental Digital Content 1, http://links.lww.com/CCM/D222).

The RIPD score has only been evaluated as a risk adjustment tool for in-hospital mortality. Some of the more mature risk adjustment algorithms such as APACHE also can be used to risk adjust for other outcomes such as ICU length of stay or time on ventilator. Evaluating whether machine learning-based methods could be used to model these outcomes would be an interesting area of investigation. In addition, we were not able to make direct comparison of the RIPD score's performance to the performance of other risk adjustment algorithms, instead we compared the observed performance of the RIPD score to previously published performance statistics for these other risk adjustment algorithms. It is also possible that some features our data, for example, the fact that the train and validation data were drawn from a common EHR system, or the inclusion of in-bound transfer patients could explain a portion of the difference in performance.

The relative influence ascribed to each feature in the model (Table 2) may be specific to our network of hospitals. Despite the demonstrable patient heterogeneity across our hospitals (Table 1), and design of our development and testing framework, further external and independent evaluations are necessary. Existing mortality risk adjustment algorithms have been developed over the course of decades and have been externally validated in a number of different settings. External evaluations, particularly those conducted using data from different EHR systems, would be important for validating whether or not a level of performance similar to that which was observed in this study could also be achieved at independent sites. These evaluations could also provide valuable perspective on the feasibility of implementing the RIPD score in new and diverse data and technology environments. We also acknowledge that local data mapping and implementation would be required to ensure that the data at an individual site match the specifications of the algorithm, and that these efforts would come at some cost to the adopting organization. A key part of the external evaluations would be to explore the feasibility and cost of these mapping and implementation efforts.

While there is first a need for external validation studies, eventually there would also need to be a single entity or community responsible for maintaining and distributing the RIPD score algorithm. We believe that this article and its appendices provide sufficient guidance on how one would go about developing a comparable mortality risk adjustment algorithm; however, having multiple versions of an algorithm would undermine one of the premises of this study, that is, that a standard risk adjustment method is needed to facilitate comparisons across a broad set of hospitals. One potential option for widespread distribution of the RIPD score would be to encapsulate the code and corresponding data requirements, definitions, and documentation in an R package. An R package is a collection of R functions, data, and compiled code in a well-defined format. R packages are freely available Comprehensive R Archive Network, a world-wide network of ftp and web servers that store identical, up-to-date, versions of code and documentation for R (https://cran.r-project.org/). This would enable hospital systems to have access to a current and consistent version of the RIPD score algorithm. In the interim, interested parties can contact the authors to receive provided fully functional versions of the RIPD and RIPD_reduced models as appendices, as well as R code that will enable users to prepare an input dataset that meets the specifications of the RIPD or RIPD_reduced models, and R code that will enable users to evaluate performance of the models using their own data. Readers can also interact directly with the dataset online (http://shiny.cac.queensu.ca/CritCareMed/RIPD/), where RIPD performance in specific patient subsets can be explored.

Finally, to some extent machine learning algorithms such as XGBoost do function as "black box," which is a departure from traditional mortality risk adjustment models that are typically based on strong conceptual models backed by clinical consensus. These traditional models also have the advantage of being relatively easy to describe, that is, it is possible to report coefficients for each variable in the model that can give the users an understanding of how the model works. However, there is a large and increasing body of evidence that machine learning algorithms excel and outperform traditional alternatives in contexts where data are abundant and diverse, and where there is high potential for complex interactions between variables (30). In addition, there are robust and well-validated methods (as were used in this study) to ensure that machine learning algorithms provide accurate estimates of risk, and that ensure that they will perform well when exposed to new data (31). Because of their attractive features, machine learning algorithms are supplanting heuristic, rule-based, and expert systems in many industries (32).

CONCLUSIONS

Our study demonstrates the feasibility of developing an automated ICU risk adjustment algorithm using machine learning methods using only highly available EHR and administrative data elements. We have shown that this algorithm, when applied across a large, clinically, and geographically diverse set of hospital ICUs, provides excellent discrimination and calibration. The discrimination and calibration of the RIPD score compares favorably with the most established ICU risk adjustment algorithms and offers additional attractive features that distinguish it from existing tools including no direct labor costs associated with manual data collection, no licensing fees, and reliance on broadly available data elements. Most of the focus in the literature has been on the technical details of

trying to get the "right" model (10); however, the fact that the vast majority of hospitals do not use any ICU risk adjustment indicates that removing the barriers of time and cost should be more prominent area of focus (2). The potential savings in cost and time afforded by using an automated risk adjustment model based on freely available open-source software should be encouraging to all those interested in increasing the transparency and measurement of ICU performance.

ACKNOWLEDGMENTS

We thank Richard Chiang and the many members of the NTT DATA services team who assisted us with data acquisition, Dr. Huiling Zhang for institutional support.

REFERENCES

1. Keegan MT, Gajic O, Afessa B: Severity of illness scoring systems in the intensive care unit. *Crit Care Med* 2011; 39:163–169
2. Breslow MJ, Badawi O: Severity scoring in the critically ill: Part 1–interpretation and accuracy of outcome prediction scoring systems. *Chest* 2012; 141:245–252
3. Breslow MJ, Badawi O: Severity scoring in the critically ill: part 2: Maximizing value from outcome prediction scoring systems. *Chest* 2012; 141:518–527
4. Salluh JI, Soares M: ICU severity of illness scores: APACHE, SAPS and MPM. *Curr Opin Crit Care* 2014; 20:557–565
5. Higgins TL, Teres D, Copes WS, et al: Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med* 2007; 35:827–835
6. Knaus WA, Draper EA, Wagner DP, et al: APACHE II: A severity of disease classification system. *Crit Care Med* 1985; 13:818–829
7. Le Gall JR, Loirat P, Alperovitch A, et al: A Simplified Acute Physiology Score for ICU patients. *Crit Care Med* 1984; 12:975–977
8. Kuzniewicz MW, Vasilevskis EE, Lane R, et al: Variation in ICU risk-adjusted mortality: Impact of methods of assessment and potential confounders. *Chest* 2008; 133:1319–1327
9. Kramer AA, Higgins TL, Zimmerman JE: Comparison of the Mortality Probability Admission Model III, National Quality Forum, and Acute Physiology and Chronic Health Evaluation IV hospital mortality models: Implications for national benchmarking. *Crit Care Med* 2014; 42:544–553
10. Glance LG, Dick AW, Osler TM: ICU scoring systems: After 30 years of reinventing the wheel, isn't it time to build the cart? *Crit Care Med* 2014; 42:732–734
11. Charles D, Gabriel M, Searcy T: Adoption of Electronic Health Record Systems Among US Non-Federal Acute Care Hospitals: 2008–2013. ONC Data Brief 23. Washington, DC, Office of the National Coordinator for Health Information Technology, 2015
12. Liu V, Turk BJ, Ragins AI, et al: An electronic Simplified Acute Physiology Score-based risk adjustment score for critical illness in an integrated healthcare system. *Crit Care Med* 2013; 41:41–48
13. Render ML, Deddens J, Freyberg R, et al: Veterans Affairs intensive care unit risk adjustment model: Validation, updating, recalibration. *Crit Care Med* 2008; 36:1031–1042
14. Center for Disease Control and Prevention: CDC locations and descriptions and instructions for mapping patient care locations. 2016. Available at: http://www.cdc.gov/nhsn/pdfs/pscmanual/15locationsdescriptions_current.pdf. Accessed Novermber 25, 2015
15. Nates JL, Nunnally M, Kleinpell R, et al: ICU admission, discharge, and triage guidelines: A framework to enhance clinical operations, development of institutional policies, and further research. *Crit Care Med* 2016; 44:1553–1602
16. Romano PS, Chan BK: Risk-adjusting acute myocardial infarction mortality: Are APR-DRGs the right tool? *Health Serv Res* 2000; 34:1469–1489
17. Heede KV den, Sermeus W, Diya L, et al: Adverse outcomes in Belgian acute hospitals: Retrospective analysis of the national hospital discharge dataset. *Int J Qual Health Care* 2006; 18:211–219
18. Levin MA, McCormick PJ, Lin HM, et al: Low intraoperative tidal volume ventilation with minimal PEEP is associated with increased mortality. *Br J Anaesth* 2014; 113:97–108
19. Bruijns SR, Guly HR, Bouamra O, et al: The value of traditional vital signs, shock index, and age-based markers in predicting trauma mortality. *J Trauma Acute Care Surg* 2013; 74:1432–1437
20. Chen T, Guestrin C: XGBoost: A Scalable Tree Boosting System. ArXiv Prepr ArXiv160302754, 2016. Available at: http://arxiv.org/abs/1603.02754. Accessed June 2, 2016
21. Chen T, He T: XGboost: eXtreme Gradient Boosting. R Package Version 04-2, 2015. Available at: http://cran.hafro.is/web/packages/xgboost/vignettes/xgboost.pdf. Accessed June 2, 2016
22. Paul P, Pennell ML, Lemeshow S: Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Stat Med* 2013; 32:67–80
23. Kramer AA, Zimmerman JE: Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med* 2007; 35:2052–2056
24. Halpern NA, Pastores SM: Critical care medicine in the United States 2000-2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med* 2010; 38:65–71
25. Coopersmith CM, Wunsch H, Fink MP, et al: A comparison of critical care research funding and the financial burden of critical illness in the United States. *Crit Care Med* 2012; 40:1072–1079
26. Bodenheimer T: High and rising health care costs. Part 2: Technologic innovation. *Ann Intern Med* 2005; 142:932–937
27. Halpern NA, Pastores SM, Thaler HT, et al: Changes in critical care beds and occupancy in the United States 1985-2000: Differences attributable to hospital size. *Crit Care Med* 2006; 34:2105–2112
28. Zimmerman JE, Kramer AA: A model for identifying patients who may not need intensive care unit admission. *J Crit Care* 2010; 25: 205–213
29. Baram D, Daroowalla F, Garcia R, et al: Use of the All Patient Refined-Diagnosis Related Group (APR-DRG) Risk of Mortality Score as a severity adjustor in the medical ICU. *Clin Med Circ Respirat Pulm Med* 2008; 2:19–25
30. Jordan MI, Mitchell TM: Machine learning: Trends, perspectives, and prospects. *Science* 2015; 349:255–260
31. Breiman L: Statistical Modeling: The two cultures (with comments and a rejoinder by the author). *Stat Sci* 2001; 16:199–231
32. Lewis-kraus G: The Great A.I. Awakening. *N Y Times*, 2016. Available at: https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html. Accessed January 31, 2018