Inclusion of Unstructured Clinical Text Improves Early Prediction of Death or Prolonged ICU Stay*

Gary E. Weissman, MD, MSHP[1,2,3]; Rebecca A. Hubbard, PhD[4]; Lyle H. Ungar, PhD[5]; Michael O. Harhay, PhD[2,4]; Casey S. Greene, PhD[6,7,8]; Blanca E. Himes, PhD[4,8]; Scott D. Halpern, MD, PhD[1,2,3,4]

Objectives: Early prediction of undesired outcomes among newly hospitalized patients could improve patient triage and prompt conversations about patients' goals of care. We evaluated the performance of logistic regression, gradient boosting machine, random forest, and elastic net regression models, with and without unstructured clinical text data, to predict a binary composite outcome of in-hospital death or ICU length of stay greater than or equal to 7 days using data from the first 48 hours of hospitalization.

Design: Retrospective cohort study with split sampling for model training and testing.

Setting: A single urban academic hospital.

Patients: All hospitalized patients who required ICU care at the Beth Israel Deaconess Medical Center in Boston, MA, from 2001 to 2012.

Interventions: None.

Measurements and Main Results: Among eligible 25,947 hospital admissions, we observed 5,504 (21.2%) in which patients died or had ICU length of stay greater than or equal to 7 days. The gradient boosting machine model had the highest discrimination without (area under the receiver operating characteristic curve, 0.83; 95% CI, 0.81–0.84) and with (area under the receiver operating characteristic curve, 0.89; 95% CI, 0.88–0.90) text-derived variables. Both gradient boosting machines and random forests outperformed logistic regression without text data ($p < 0.001$), whereas all models outperformed logistic regression with text data ($p < 0.02$). The inclusion of text data increased the discrimination of all four model types ($p < 0.001$). Among those models using text data, the increasing presence of terms "intubated" and "poor prognosis" were positively associated with mortality and ICU length of stay, whereas the term "extubated" was inversely associated with them.

Conclusions: Variables extracted from unstructured clinical text from the first 48 hours of hospital admission using natural language processing techniques significantly improved the abilities of logistic regression and other machine learning models to predict which patients died or had long ICU stays. Learning health systems may adapt such models using open-source approaches to capture local variation in care patterns. (*Crit Care Med* 2018; 46:1125–1132)

Key Words: critical care; decision support techniques; electronic health records; forecasting; machine learning; natural language processing

***See also p. 1196.**

[1]Division of Pulmonary, Allergy, and Critical Care, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

[2]Palliative and Advanced Illness Research Center, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

[3]Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA.

[4]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

[5]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA.

[6]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA.

[7]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

[8]Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

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).

Drs. Weissman, Ungar, and Himes received support for article research from the National Institutes of Health (NIH). Dr. Weissman was supported by NIH T32 HL098054 for this work. Dr. Ungar's institution received funding from the NIH, Defense Advanced Research Projects Agency, and Templeton Research Trust. Dr. Himes' institution received funding from the NIH/National Heart, Lung and Blood Institute. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: gary.weissman@uphs.upenn.edu

DOI: 10.1097/CCM.0000000000003148

Admission to the ICU is costly (1), associated with long-term sequelae among patients and families (2), places strain on healthcare delivery systems (3–5), and may not always reflect patient wishes (6–8). Patients admitted to the ICU commonly die (9), and others experience prolonged

ICU stays that are associated with greater costs and functional impairments following discharge (8, 10). If the risk of such adverse outcomes were identified early during a hospitalization, caregivers might have opportunities to increase preference-concordant, high-value care. For example, clinicians could initiate goals of care discussions earlier and with more certainty or could allocate staffing and other clinical resources to reduce capacity strain or streamline transitions in care (11, 12). However, ICU clinicians' predictions of these outcomes are imperfect (13–17), and prior efforts to predict mortality (9, 18, 19) or long ICU length of stay (LOS) (20–22) have had limited applicability at the bedside (9, 20, 23).

Among the factors limiting the clinical use of prior models to predict adverse hospitalization outcomes are the use of structured data fields alone, the nearly universal use of logistic regression models which assume linear and additive relationships among independent variables, and the lack of replicable statistical code, thereby limiting the ability to customize models to individual hospitals or health systems (6, 24). We therefore sought to overcome each of these three limitations. We were motivated, first, by the potential for natural language processing to identify features of critical illness among ICU patients in progress notes (25) and in discharge summaries (26) that may not be found in structured data fields (27, 28); second, by observations that machine learning models may provide superior discrimination of mortality predictions compared with traditional regression models (29–32); and third, by the virtues of producing models that can be assessed for reproducibility in other settings (33). Previous work has also demonstrated increases in the discrimination of ICU mortality prediction models using text data from nursing notes (34, 35).

Accordingly, we sought first to build clinical prediction models that would identify, early in the course of hospitalization, a composite outcome of in-hospital death or ICU LOS greater than or equal to 7 days; second, to describe the predictive performance of such models built using both traditional regression and machine learning approaches; and third, to complete these objectives using an easily reproducible, open-source workflow in accordance with recommendations for electronic health record (EHR)–based predictive analytics (36). We hypothesized that the inclusion of data derived from the unstructured text of clinical encounter notes would improve the performance of these prediction models.

MATERIALS AND METHODS

Study Population

We analyzed the Medical Information Mart for Intensive Care (MIMIC) III, a publicly available, deidentified dataset that includes information from all hospitalizations requiring ICU care at the Beth Israel Deaconess Medical Center in Boston, MA, from 2001 to 2012 (37). We excluded patients less than 18 years old, admissions without any clinical encounter notes recorded in the first 48 hours of hospital admission (regardless of how many of those hours were spent in an ICU), admissions with total hospital LOS less than 48 hours. Those with a documented limitation on life-sustaining therapy within the first 48 hours were also excluded to improve the relevance of the predictions. This study was considered exempt by the Institutional Review Board of the University of Pennsylvania.

Model Development and Comparison

Four different modeling approaches (**Supplemental table e5**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511) were selected to compare different variable selection methods and relationships between predictor variables (for descriptions, see **Digital Supplement**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511). The selected model types are similar to those in previous studies that demonstrated machine learning approaches with nonlinear decision boundaries may improve performance in a population with critical illness (38). For each approach, we first built a model using only variables from structured data fields and then built an identical model that also included variables derived from unstructured data.

To train and evaluate models, we divided our dataset randomly into a training sample, consisting of 75% of encounters, and a testing sample, consisting of the remaining 25% of encounters which were withheld from all models during the training process. Tuning variables for each model were estimated using five times repeated, 10-fold cross-validation on the training sample (**Supplemental fig. e1**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511). Once the optimal tuning variables were determined, the final models were fitted with the full training sample. Model performance measures were reported using the testing sample. Model error was minimized with respect to the area under the receiver operating characteristic curve (AUC). A full working code example that contains details of the model building and text processing approaches is available online (39).

Model discrimination was assessed with the AUC. Comparisons between models were made with the DeLong test (40) using a Bonferroni correction to the family-wise error rate for multiple comparisons ($m = 28$) in the primary analysis. Given the large sample size, model calibration was assessed via visual inspection of the calibration curves (41). Comparisons between demographic and clinical subgroups were made with chi-square and Wilcoxon rank-sum tests for categorical and continuous measures, respectively.

Structured Data Sources

All models included 18 variables commonly found in the EHR or used in other ICU mortality and LOS prediction models (9, 18–22, 29). These included laboratory values, vital signs, clinical events, demographics, and comorbidity burden (**Supplemental methods**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511).

Unstructured Data Sources

All clinical text notes from physicians, nurses, and other clinicians time-stamped during the first 48 hours of the hospitalization were aggregated into a single document for each

admission. Using a "closed vocabulary" approach (42), we explicitly searched for 21 key terms that we determined a priori to be relevant to identifying patients at risk for the primary outcome (**Supplemental table e1**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511). Using an "open vocabulary" approach (42, 43), we also built a document-term matrix of one-, two-, and three-word phrases (i.e., n-grams) appearing in at least 5% of all documents in the training set and used penalized logistic regression with 10-fold cross-validation to identify the 500 most predictive phrases (**Supplemental table e2**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511). This variable selection was conducted in the training set only. Using a "bag of words" approach, the integer counts of phrases appearing in the aggregated clinical notes were included as continuous variables in the set of models using unstructured text data (44). The total word count of each aggregated set of documents was also included as a covariate in the models using text data.

Primary Outcome and Analysis

The primary outcome for the main analysis was a composite of in-hospital death or ICU LOS greater than or equal to 7 days. We chose this composite outcome because LOS cannot be assessed validly without considering mortality at the same time (45) and because both mortality and prolonged ICU LOS represent unfavorable outcomes that might prompt early decisions around clinical care or hospital resource allocation.

Secondary Analyses

We conducted six secondary analyses to test the robustness and potential applicability in other settings of our approach using different outcomes and time horizons. Specifically, we built a full set of models 1) using only predictor data from the first 24 hours of hospitalization, 2) using in-hospital death as the primary outcome, 3) using in-hospital death as the outcome and only predictor data from the first 24 hours, 4) using a composite outcome of in-hospital death or ICU LOS greater than or equal to 21 days, and 5) without baseline comorbidity data. This last analysis was chosen because comorbidity burden, as measured by the Elixhauser score (46), is based on diagnostic codes determined retrospectively at the end of a hospitalization, and thus its inclusion may result in overly optimistic predictive performance since some diagnoses may not have been known within the first 48 hours of admission. Finally, to assess the contribution of the large number of predictor variables, we also 6) built a more parsimonious model with the top performing model type using only the 25 most predictive variables.

Variable Importance

The relative variable importance for each model was estimated using a type-specific approach (47). Because not all the chosen model types yield results equivalent to a β variable with a CI from a logistic regression model, the variable importance measures are not directly comparable across types. Therefore, we report the mean relative variable importance across all models. For comparison, we also report the most predictive variables from the logistic regression model (defined as those with $p < 0.05$ ranked by odds ratio [OR)] or 1/OR for OR < 1).

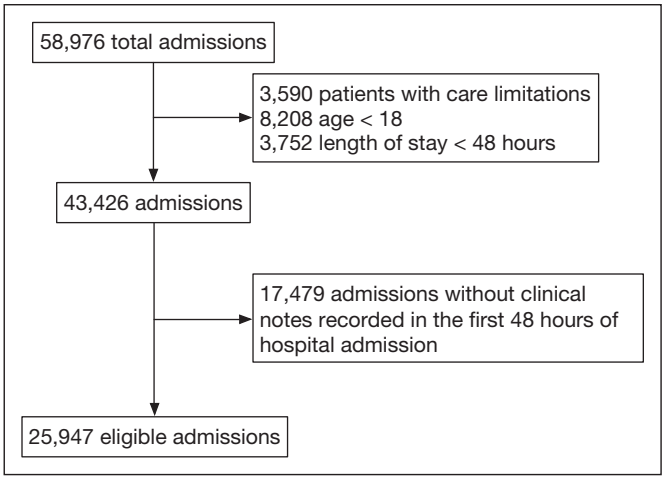

Figure 1. Study cohort and exclusions.

RESULTS

There were 58,976 hospital admissions in the dataset, of which 43,426 (73.6%) met clinical criteria for eligibility and 25,947 of these (59.7%) had complete documentation required for the analyses (**Fig. 1**). Compared with other clinically eligible patients, those excluded due to absence of clinical encounter notes were more likely to originate from "clinic referral/premature" admissions (42.4% vs 5.5%; $p < 0.001$) rather than an emergency department (13.1% vs 58.4%; $p < 0.001$), had a higher median Elixhauser score (11 vs 8; $p < 0.001$), but demonstrated no clinically relevant differences in the primary outcome (21.9% vs 21.2%; $p < 0.095$) or ICU LOS (median 2.2 vs 2.5 d; $p < 0.001$). Among the eligible sample, the median hospital LOS was 6.9 days (interquartile range [IQR], 4.5–11.6 d), the median ICU LOS was 2.5 days (IQR, 1.4–5.0 d), and 5,504 (21.2%) either died or had ICU LOS greater than or equal to 7 days (**Supplemental figs. e11** and **e12**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511). The study population is further described in **Table 1**. Laboratory values and other inputs are reported in **Supplemental table e3** (Supplemental Digital Content 1, http://links.lww.com/CCM/D511) and missingness in **Supplemental table e4** (Supplemental Digital Content 1, http://links.lww.com/CCM/D511).

Model Performance

Among the four model types trained, the gradient boosting machine model had the highest discrimination without unstructured text data (AUC, 0.83; 95% CI, 0.81–0.84) and with such data (AUC, 0.89; 95% CI, 0.88–0.90). The addition of unstructured text data improved the performance of all models ($p < 0.001$) (**Fig. 2**). Performance results from both testing and training samples are found in **Table 2** (see also **Supplemental figs. e5-e10**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511). Among models that included unstructured text data, all performed better than logistic regression ($p < 0.02$). All models had at least fair calibration by visual inspection (**Fig. 3**).

TABLE 1. Characteristics of the Study Population Stratified by Whether They Experienced the Primary Outcome of In-Hospital Death or ICU Length of Stay Greater Than or Equal to 7 Days

Characteristics	All	Died or ICU LOS ≥ 7 d	
		Yes	No
Patient admissions, *n* (%)	25,947	5,504 (21.2)	20,443 (78.8)
In-hospital death, *n* (%)	1,861 (7.8)	1,861 (33.8)	0 (0)
ICU LOS ≥ 7 d, *n* (%)	4,564 (17.6)	4,564 (82.9)	0 (0)
Age (yr), median (IQR)	64.1 (51.3–76.3)	66.9 (53.6–77.8)	63.4 (50.1–75.8)
ICU LOS (d), median (IQR)	2.5 (1.4–5.0)	11.0 (7.8–17.6)	2.0 (1.2–3.2)
Hospital LOS (d), median (IQR)	6.9 (4.5–11.6)	16.5 (10.7–25.0)	6.0 (4.2–8.5)
Male, *n* (%)	14,958 (57.5)	3,161 (57.4)	11,797 (57.7)
Admission type, *n* (%)			
Emergency	20,867 (80.4)	4,837 (87.9)	16,030 (78.4)
Elective	4,341 (16.7)	451 (8.2)	3,890 (19.0)
Urgent	739 (2.8)	216 (3.9)	523 (2.6)
Initial ICU type, *n* (%)			
Medical ICU	9,622 (37.1)	2,168 (39.4)	7,454 (36.5)
Cardiothoracic surgery care unit	4,935 (19.0)	686 (12.5)	4,249 (20.8)
Cardiac care unit	4,030 (15.5)	765 (13.9)	3,265 (16.0)
Surgical ICU	4,024 (15.5)	1,031 (18.7)	2,993 (14.6)
Trauma surgery ICU	3,329 (12.8)	853 (15.5)	2,476 (12.1)
Not recorded	7 (< 0.1)	1 (< 0.1)	6 (< 0.1)
Self-reported race or ethnicity, *n* (%)			
White	18,461 (71.1)	3,916 (71.4)	14,545 (71.1)
Black	2,475 (9.5)	407 (7.4)	2,068 (10.1)
Hispanic or Latino	848 (3.3)	151(2.7)	697 (3.4)
Asian	563 (2.2)	97 (1.8)	466 (2.3)
Other/unknown	3,600 (13.9)	933 (17.0)	2,667 (13.0)
Modified Elixhauser score, median (IQR)	8 (2–15)	12 (6–19)	6 (1–13)
Clinical deterioration, *n* (%)			
Mechanical ventilation	4,512 (17.4)	1,926 (35.0)	2,586 (12.6)
Cardiac arrest	206 (0.8)	91 (1.7)	115 (0.6)
ICU transfer	25,614 (98.7)	5,427 (98.6)	20,187 (98.7)
Laboratory data, median (IQR)			
Creatinine, highest	1.10 (0.80–1.60)	1.20 (0.90–2.10)	1.00 (0.80–1.50)
WBC count, highest	13.0 (9.70–17.10)	14.8 (10.9–19.8)	12.6 (9.5–16.5)
Platelets, lowest	176 (125–235)	164 (106–228)	179 (130–237)
Total bilirubin, highest	1.5 (0.6–1.5)	1.5 (0.6–1.5)	1.5 (0.6–1.5)
Pao_2, lowest	103 (93–103)	103 (71–103)	103 (103–103)
Potassium, highest	4.5 (4.2–4.9)	4.6 (4.2–5.1)	4.5 (4.1–4.9)

(*Continued*)

TABLE 1. (*Continued*). Characteristics of the Study Population Stratified by Whether They Experienced the Primary Outcome of In-Hospital Death or ICU Length of Stay Greater Than or Equal to 7 Days

Characteristics	All	Died or ICU LOS ≥ 7 d	
		Yes	No
Vital signs, median (IQR)			
Urine output (cc/kg/hr)	0.85 (0.57–0.90)	0.85 (0.53–0.91)	0.85 (0.58–0.90)
Glasgow Coma Scale, lowest	9 (7–14)	8 (3–9)	9 (9–15)
Systolic blood pressure (mm Hg), lowest	89 (79–101)	83 (70–94)	91 (81–102)
Heart rate (beats/min), highest	69 (60–78)	69 (60–80)	68 (60–77)
Temperature (°C), highest	36.0 (35.6–36.4)	35.9 (35.4–36.4)	36.1 (35.6–36.4)
Clinical notes, median (IQR)			
Total raw word count	1,023 (562–2,266)	1,410 (956–2,292)	899 (501–2,247)
Number of notes	6 (4–11)	8 (6–12)	5 (3–10)

IQR = interquartile range, LOS = length of stay.
All clinical variables were measured during the first 48 hr of the hospital admission.

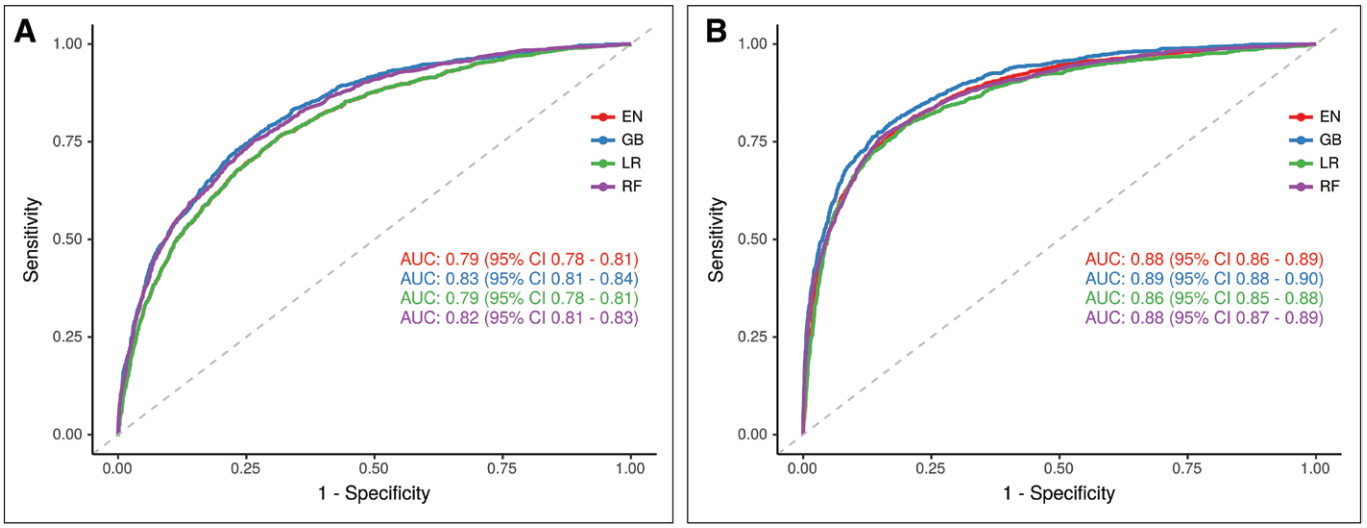

Figure 2. Receiver operating characteristic curve of models using only structured (**A**) and both structured and unstructured (**B**) data sources. AUC = area under the receiver operating characteristic curve, EN = elastic net, GB = gradient boosting machines, LR = logistic regression, RF = random forest.

TABLE 2. Area Under the Receiver Operating Characteristic Curve for Each Model Type, With and Without Unstructured Text Data, and Using the Training (75%) and Testing (25%) Samples

Model Types	Unstructured Data Only (AUC)		Structured and Unstructured Data (AUC)	
	Training	Testing	Training	Testing
Logistic regression	0.80 (0.79–0.81)	0.79 (0.78–0.81)	0.90 (0.89–0.90)	0.86 (0.85–0.88)
Elastic net regression	0.80 (0.79–0.81)	0.79 (0.78–0.81)	0.89 (0.88–0.89)	0.88 (0.86–0.89)
Random forests	1.00 (1.00–1.00)	0.82 (0.81–0.83)	1.00 (1.00–1.00)	0.88 (0.87–0.89)
Gradient boosting machines	0.89 (0.89–0.90)	0.83 (0.81–0.84)	0.95 (0.95–0.96)	0.89 (0.88–0.90)

AUC = area under the receiver operating characteristic curve.

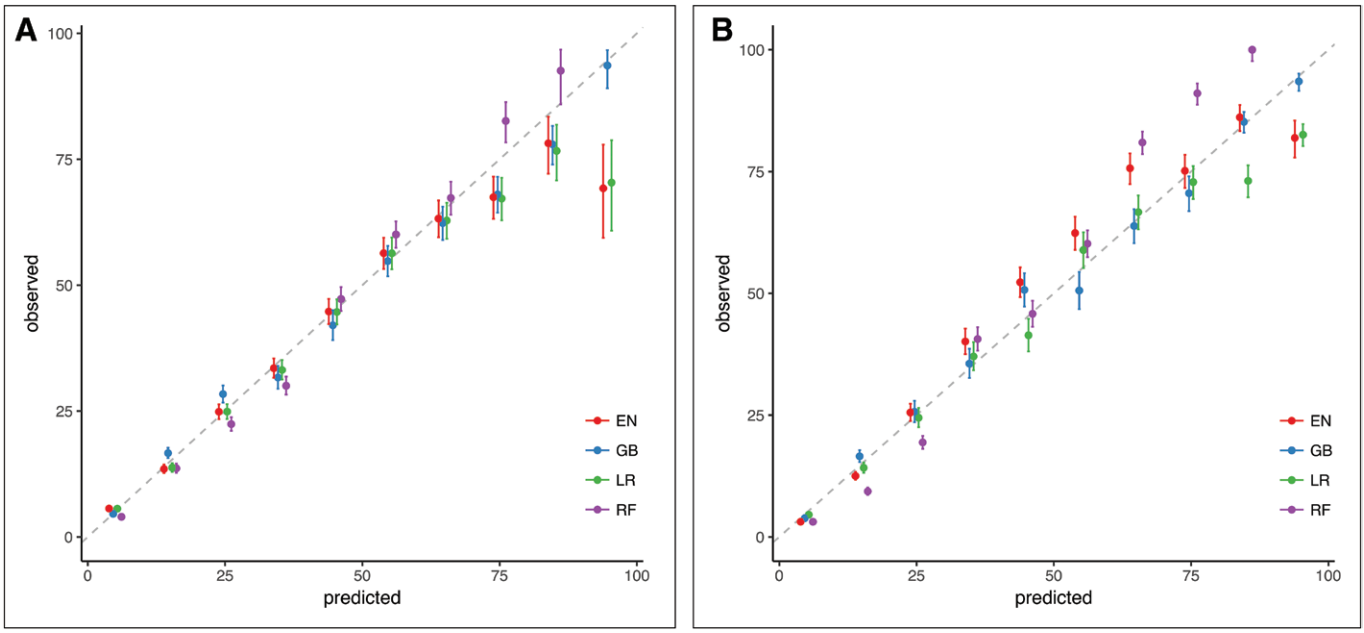

Figure 3. Calibration plot of models using only structured (**A**) and both structured and unstructured (**B**) data sources. EN = elastic net, GB = gradient boosting machines, LR = logistic regression, RF = random forest.

The gradient boosting machine model with unstructured text data performed slightly better among patients less than 65 years old (AUC, 0.91 vs 0.87; $p < 0.001$). The discrimination was equivalent between those who did or did not receive mechanical ventilation in the first 48 hours of hospitalization (AUC, 0.88 vs 0.88; $p < 0.964$) and between those first admitted to medical or surgical ICUs (AUC, 0.89 vs 0.90; $p < 0.299$).

Secondary Analyses

In secondary analyses (Digital Supplement, Supplemental Digital Content 1, http://links.lww.com/CCM/D511), model discrimination also increased in all model types with the inclusion of unstructured text data. The restriction to an informative time horizon of 24 hours yielded slightly lower AUC point estimates, but significant improvements were observed with the use of unstructured data for all models over structured data alone ($p < 0.001$). When the Elixhauser comorbidity data were removed, the AUCs were also lower, but the increase in performance with the addition of unstructured text data persisted across all models ($p < 0.001$).

The gradient boosting machine model restricted to the 25 most predictive covariates had similar performance (AUC, 0.88; 95% CI, 0.87–0.89) to that in the primary analysis (**Supplemental figs. e3** and **e4**, http://links.lww.com/CCM/D511). Detailed results from the secondary analyses are found in **Supplemental tables e6-e8** (Supplemental Digital Content 1, http://links.lww.com/CCM/D511).

Variable Importance

Extraction of text features yielded 5,790 unique phrases from which 500 were chosen for inclusion based on penalized regression (Supplemental table e1, Supplemental Digital Content 1, http://links.lww.com/CCM/D511). In the logistic regression model using only structured data, the three most predictive variables were urgent (OR, 4.78; 95% CI, 3.75–6.10) and emergency (OR, 3.32; 95% CI, 2.92–3.78) admission types and any mechanical ventilation in the first 48 hours (OR, 3.11; 95% CI, 2.84–3.77). In contrast, in the logistic regression model also using structured data, the most predictive variables were the term "poor prognosis" (OR, 3.04; 95% CI, 1.71–5.40), urgent admission (OR, 3.00; 95% CI, 1.71–5.40), and any ICU transfer in the first 48 hours (OR, 0.44; 95% CI, 0.33–0.62). Across all models, seven of the ten most predictive variables were the words and phrases derived from the unstructured text data (**Supplemental fig. e2**, Supplemental Digital Content 1, http://links.lww.com/CCM/D511).

DISCUSSION

This study shows that variables derived from the unstructured text of clinical notes significantly improved the discrimination of models designed to predict in-hospital death or prolonged ICU LOS within the first 24 or 48 hours of a hospitalization. These improvements were apparent across all model types, including those using traditional regression and other machine learning approaches. These results should not be construed to mean that the individual models trained here should be used in other health systems or ICUs. Rather, they indicate several ways in which health systems may develop and use clinical prediction models based on available free-text data in the EHR to improve care for their patients.

First, the significant increase in model discrimination using text data suggests a rich opportunity to improve the performance of health system–level clinical prediction models. Existing ICU mortality and LOS models, trained on large, national datasets, are limited in that they do not have access to and therefore cannot rely upon the text of clinical notes. Seven of the 10 most predictive variables in models using text data were derived from unstructured clinical text. Future

development of clinical prediction models for patients in the ICU should leverage this information-rich resource.

Second, performance of the gradient boosting machine model—the top performing model in the primary analysis—remained high across patient age, ICU type, and need for mechanical ventilation despite clinical and demographic heterogeneity in those populations. Locally trained models with such high discrimination that are robust across multiple subgroups are more likely to be useful at the bedside.

Third, some of the machine learning models, especially tree-based models such as gradient boosting machines and random forests, outperformed logistic regression with or without the inclusion of unstructured data. These modeling approaches are suited to capturing nonlinear and otherwise complex relationships among many predictor variables and may be particularly amenable to modeling the complexity of health states found in patients with critical illness. The best model in this study compared at least as well to performance in prior work that used a machine learning–based mortality model with an older version of the MIMIC dataset (29). Similarly, although several other models have been developed to predict mortality among ICU patients with AUCs of 0.88 (9), 0.848 (18), and 0.823 (19) and to predict ICU LOS as a continuous outcome with R^2 values of 0.215 (21) and 0.202 (22), these models are not directly comparable with those in our primary analysis because of differences in informative time horizons and outcomes. Although our preliminary work included all patients regardless of care limitations and produced models with even better discrimination (48), the present study maintains high performance in a population more likely to benefit from bedside prognostic models. Even in this restricted population, our mortality model in the secondary analysis compares about as well or better than the aforementioned mortality models. The nearly equivalent performance, however, of the parsimonious model with only 25 variables reveals opportunities to reduce overfitting in future model development, the need to examine other variable relationships not explored in this study, and supports investigation of other data sources in order to improve performance of mortality and LOS prediction models.

Fourth, by using open-source methods and including the code we used in the analysis (39), we have provided a reproducible workflow that health systems can use to rapidly build accurate and customized versions of these prediction models to ensure similar or perhaps greater accuracy in local settings. Future work that employs transparent and reproducible methods will increase opportunities for collaboration to modify and improve upon these approaches.

This study has several strengths. First, it relies on traditional administrative and clinical data sources in addition to the rich text of clinical encounter notes. This combination of inputs reflects the data available to modern health systems in an era of widespread EHR adoption (49). Second, the use of different regression and machine learning model types mitigates limitations in prior work of model misspecification bias. Finally, this study provides a key stepping stone for health systems to begin developing their own locally customized clinical prediction models for patients admitted to an ICU.

This study also has limitations. First, the generalizability of the findings is limited by the dataset, which was preselected to include only patients who had an ICU stay during their hospital admission and was collected from a single, urban academic hospital up until 2012. However, the general principles revealed about ways to improve clinical prediction among critically ill patients are likely to generalize beyond this setting. Second, it is possible that patients excluded for having only radiology notes rather than clinical encounter notes during the first 48 hours of admission would encounter differences in triage, care, and documentation practices, potentially further limiting the generalizability of the trained models to all patients who are admitted to hospitals. Third, we did not assess the relative improvement in prognostication that our models might yield relative to how "obvious" a prognosis might be to a clinician. Finally, we did not address the potential for self-reinforcing bias in clinician attitudes and potential responses to a predictive algorithm based on biased text.

CONCLUSIONS

Free-text data from clinical notes can significantly improve the accuracy of models that predict in-hospital mortality and prolonged ICU LOS within the first 24 or 48 hours of hospital admission. Machine learning approaches can produce very accurate clinical prediction models and may be superior to traditional logistic regression models when using large numbers of predictor variables with varying relationships. These improvements may be due in part to the ability of such modeling approaches to capture nonlinear decision boundaries in complex patients with critical illness. Statistical workflows that use open-source software to generate these models can be disseminated transparently for local adaption by learning health systems.

REFERENCES

1. Barrett M, Smith M, Elixhauser A, et al: Utilization of Intensive Care Services, 2011. Technical Report #185. Rockville, Agency for Healthcare Research and Quality, 2014
2. Elliott D, Davidson JE, Harvey MA, et al: Exploring the scope of post-intensive care syndrome therapy and care: Engagement of non-critical care providers and survivors in a second stakeholders meeting. *Crit Care Med* 2014; 42:2518–2526
3. Gabler NB, Ratcliffe SJ, Wagner J, et al: Mortality among patients admitted to strained intensive care units. *Am J Respir Crit Care Med* 2013; 188:800–806
4. Wagner J, Gabler NB, Ratcliffe SJ, et al: Outcomes among patients discharged from busy intensive care units. *Ann Intern Med* 2013; 159:447–455
5. Weissman GE, Gabler NB, Brown SE, et al: Intensive care unit capacity strain and adherence to prophylaxis guidelines. *J Crit Care* 2015; 30:1303–1309
6. Hart JL, Harhay MO, Gabler NB, et al: Variability among US intensive care units in managing the care of patients admitted with pre-existing limits on life-sustaining therapies. *JAMA Intern Med* 2015; 175:1019–1026
7. Le Guen J, Boumendil A, Guidet B, et al: Are elderly patients' opinions sought before admission to an intensive care unit? Results of the ICE-CUB study. *Age Ageing* 2016; 45:303–309
8. Teno JM, Fisher E, Hamel MB, et al: Decision-making and outcomes of prolonged ICU stays in seriously ill patients. *J Am Geriatr Soc* 2000; 48:S70–S74

9. Zimmerman JE, Kramer AA, McNair DS, et al: Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34:1297–1310
10. Fan E, Dowdy DW, Colantuoni E, et al: Physical complications in acute lung injury survivors: A two-year longitudinal prospective study. *Crit Care Med* 2014; 42:849–859
11. Jones SS, Thomas A, Evans RS, et al: Forecasting daily patient volumes in the emergency department. *Acad Emerg Med* 2008; 15:159–170
12. Kerlin MP, Harhay MO, Vranas KC, et al: Objective factors associated with physicians' and nurses' perceptions of intensive care unit capacity strain. *Ann Am Thorac Soc* 2014; 11:167–172
13. Nassar AP Jr, Caruso P: ICU physicians are unable to accurately predict length of stay at admission: A prospective study. *Int J Qual Health Care* 2016; 28:99–103
14. Rocker G, Cook D, Sjokvist P, et al; Level of Care Study Investigators; Canadian Critical Care Trials Group: Clinician predictions of intensive care unit mortality. *Crit Care Med* 2004; 32:1149–1154
15. Meadow W, Pohlman A, Frain L, et al: Power and limitations of daily prognostications of death in the medical intensive care unit. *Crit Care Med* 2011; 39:474–479
16. Sinuff T, Adhikari NK, Cook DJ, et al: Mortality predictions in the intensive care unit: Comparing physicians with scoring systems. *Crit Care Med* 2006; 34:878–885
17. Detsky ME, Harhay MO, Bayard DF, et al: Discriminative Accuracy of physician and nurse predictions for survival and functional outcomes 6 months after an ICU admission. *JAMA* 2017; 317:2187–2195
18. Moreno RP, Metnitz PG, Almeida E, et al; SAPS 3 Investigators: SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; 31:1345–1355
19. 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
20. Verburg IW, Atashi A, Eslami S, et al: Which models can I use to predict adult ICU length of stay? A systematic review. *Crit Care Med* 2017; 45:e222–e231
21. Zimmerman JE, Kramer AA, McNair DS, et al: Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med* 2006; 34:2517–2529
22. Kramer AA, Zimmerman JE: A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. *BMC Med Inform Decis Mak* 2010; 10:27
23. Kramer AA: Are ICU length of stay predictions worthwhile? *Crit Care Med* 2017; 45:379–380
24. Admon AJ, Seymour CW, Gershengorn HB, et al: Hospital-level variation in ICU admission and critical care procedures for patients hospitalized for pulmonary embolism. *Chest* 2014; 146:1452–1461
25. Abhyankar S, Demner-Fushman D, Callaghan FM, et al: Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *J Am Med Inform Assoc* 2014; 21:801–807
26. Weissman GE, Harhay MO, Lugo RM, et al: Natural language processing to assess documentation of features of critical illness in discharge documents of acute respiratory distress syndrome survivors. *Ann Am Thorac Soc* 2016; 13:1538–1545
27. Ford E, Carroll JA, Smith HE, et al: Extracting information from the text of electronic medical records to improve case detection: A systematic review. *J Am Med Inform Assoc* 2016; 23:1007–1015
28. Navathe AS, Zhong F, Lei VJ, et al: Hospital readmission and social risk factors identified from physician notes. *Health Serv Res* 2018; 53:1110–1136
29. Pirracchio R, Petersen ML, Carone M, et al: Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *Lancet Respir Med* 2015; 3:42–52
30. Le Gall JR, Lemeshow S, Saulnier F: A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; 270:2957–2963
31. LaFaro RJ, Pothula S, Kubal KP, et al: Neural network prediction of ICU length of stay following cardiac surgery based on pre-incision variables. *PLoS One* 2015; 10:e0145395
32. Tsai PFJ, Chen PC, Chen YY, et al: Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. *J Healthc Eng* 2016; 2016
33. Goldstein BA, Navar AM, Pencina MJ, et al: Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J Am Med Inform Assoc* 2017; 24:198–208
34. Marafino BJ, Boscardin WJ, Dudley RA: Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *J Biomed Inform* 2015; 54:114–120
35. Lehman LW, Saeed M, Long W, et al: Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp Proc* 2012; 2012:505–511
36. Amarasingham R, Audet AM, Bates DW, et al: Consensus statement on electronic health predictive analytics: A guiding framework to address challenges. *EGEMS (Wash DC)* 2016; 4:1163
37. Johnson AE, Pollard TJ, Shen L, et al: MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3:160035
38. Churpek MM, Yuen TC, Winslow C, et al: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016; 44:368–374
39. Weissman GE: gweissman/text-pred-icu: Full release, 2017. Available at: https://doi.org/10.5281/zenodo.826438. Accessed November 1, 2017
40. DeLong ER, DeLong DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; 44:837–845
41. 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
42. Kern ML, Park G, Eichstaedt JC, et al: Gaining insights from social media language: Methodologies and challenges. *Psychol Methods* 2016; 21:507–525
43. Schwartz HA, Eichstaedt JC, Kern ML, et al: Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One* 2013; 8:e73791
44. Salton G, McGill M: Introduction to Modern Information Retrieval, New York, NY, McGraw-Hill, 1983
45. Harhay MO, Ratcliffe SJ, Halpern SD: Measurement error in intensive care unit length of stay estimates due to patient flow. *Am J Epidemiol* 2017; 186:1389–1395
46. Elixhauser A, Steiner C, Harris DR, et al: Comorbidity measures for use with administrative data. *Med Care* 1998; 36:8–27
47. Kuhn M: caret: Variable Importance, 2017. Available at: https://topepo.github.io/caret/variable-importance.html. Accessed November 1, 2017
48. Weissman GE, Hubbard RA, Ungar LH, et al: Inclusion of unstructured text data from clinical notes improves early prediction of death or prolonged ICU stay among hospitalized patients. Poster Presentation. *Am J Respir Crit Care Med* 2017; 195:A1084
49. Charles D, Gabriel M, Searcy T: Adoption of electronic health record systems among U.S. non - federal acute care hospitals: 2008–2014. *ONC Data Brief* 2015; 23:1–10