Digital Supplement: Inclusion of Unstructured Clinical Text Improves Early Prediction of Death or Prolonged ICU Stay

Critical Care Medicine, 2018

Gary E. Weissman, Rebecca A. Hubbard, Lyle H. Ungar, Michael O. Harhay, Casey S. Greene, Blanca E. Himes, Scott D. Halpern

Supplemental Methods

Machine Learning Methods

Logistic regression

We developed two types of logistic regression models in this study. First, a standard model was trained with predictor variable inputs as described in the main analysis. Second, we trained an elastic net model that "penalizes" complex models with non-zero coefficients, thus reducing the number of variables in a final model (1). A grid search determined the optimal combination of the mixing and regularization parameters for the elastic net model.

Tree-based models

We developed two types of tree-based models in this study. First, we fit a random forest model which generates a weighted prediction based on predictive performance across many smaller trees using repeatedly sampled random subsets of variables (2). A grid search determined the optimal number of variables to consider at each node in the tree. Second, we trained a gradient boosting machine model that trains a series of trees where each subsequent tree attempts to correct the errors of the prior trees (3). For this model, a grid search determined the optimal number of trees, the number of splits to perform on each tree, the regularization parameter, and the minimum number of observations in each terminal node.

Structured Data Sources

Models included continuous variables for age (years), urine output (cc/kg/hr), and modified Elixhauser score (4–7), and the most severe value within the first 48 hours of hospital admission of creatinine (mg/dL; highest), white blood cell count (K/ul; highest), platelet count (K/ul; lowest), total bilirubin (mg/dL; highest), arterial partial pressure of oxygen (mmHg; lowest), Glasgow coma scale (lowest), systolic blood pressure (mmHg; lowest), heart rate (beats per minute; highest), and body temperature (Celsius; highest) were included as continuous variables. Categorical variables included gender, admission type (emergency, elective, or urgent), and the presence in the nursing flowsheet within the first 48 hours of mechanical ventilation, ICU admission, and cardiac arrest.

Imputation of missing vitals and lab values was performed by replacing them with the mean of non-missing values (8) rather than excluding them entirely (9) to preserve as many subjects as possible while not biasing the distribution of these variables toward extreme values.

Programming languages and libraries

Data cleaning and visualization tasks were performed with the *data.table* (10) and *ggplot2* (11) packages, respectively. Model training and validation were conducted using the *caret* package (12) with the R language for statistical computing (version 3.3.2) (13). The processing and extraction of variables from the text of clinical notes used the *Natural Language Toolkit* (14) and *scikit-learn* (15) modules for the Python programming language (version 3.5.1).

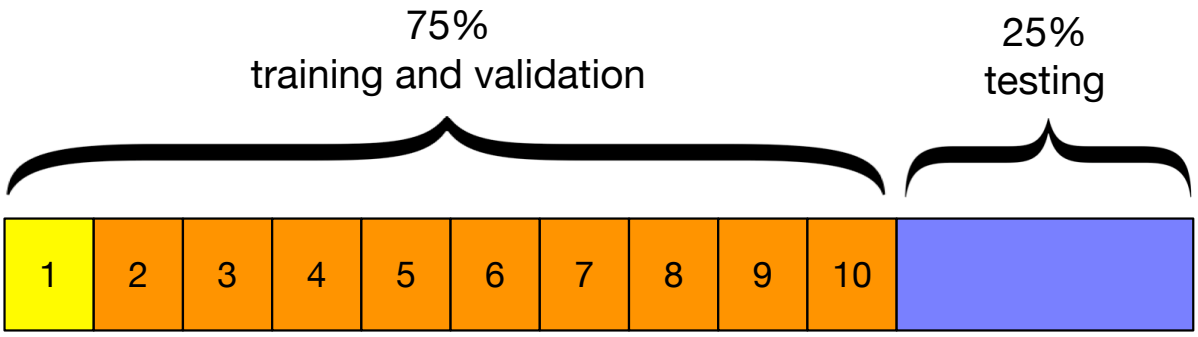
Unstructured Data Sources and Natural Language Processing Tasks

Natural language processing (NLP) techniques encompass a broad variety of tasks that researchers might use to extract relevant information from clinical text (18). First, the use of medical abbreviations are common, and if not counted, may miss the fact that "PND" is another use of "paroxysmal nocturnal dyspnea". However, if not also disambiguated, may miss that "PND" could also mean "post-nasal drip" in a different context. Although attempts have been made to overcome these abbreviation expansion and disambiguation issues, no gold standard exists for this task (19). Second, the copy-and-paste problem is also prevalent in the use of electronic health records (20), and potentially biases the identification of important keywords towards older clinical states. The dataset used in this analysis did not contain meta-data on copy/paste origin of text and we were not able to analyze the effect of using such data. Third, the use of relation extraction methods to identify temporal modifiers to terms (e.g. "she used to have chest pain" vs "she has chest pain") is necessary for understanding the contributions of concepts to an analysis of the present clinical situation. This is a very complex task (21) and was not addressed in this study. Fourth, orthographic (i.e. spelling) errors or variations may also affect the identification of identical terms (e.g. "respiratory" and "respiratroy"). In our previous work (22) in which we used a string similarity matching algorithm to link such similar terms, we found no difference in the overall rates of identification of important concepts, and so this approach was not employed in the present analysis.

Our study did not employ the foregoing NLP tasks. Even so, our approach using n-grams with a penalized regression for feature selection still resulted in significant improvements in the discrimination of the predictive models. This suggests that future work that does incorporate these NLP tasks may demonstrate even greater improvements over models using structured data alone.

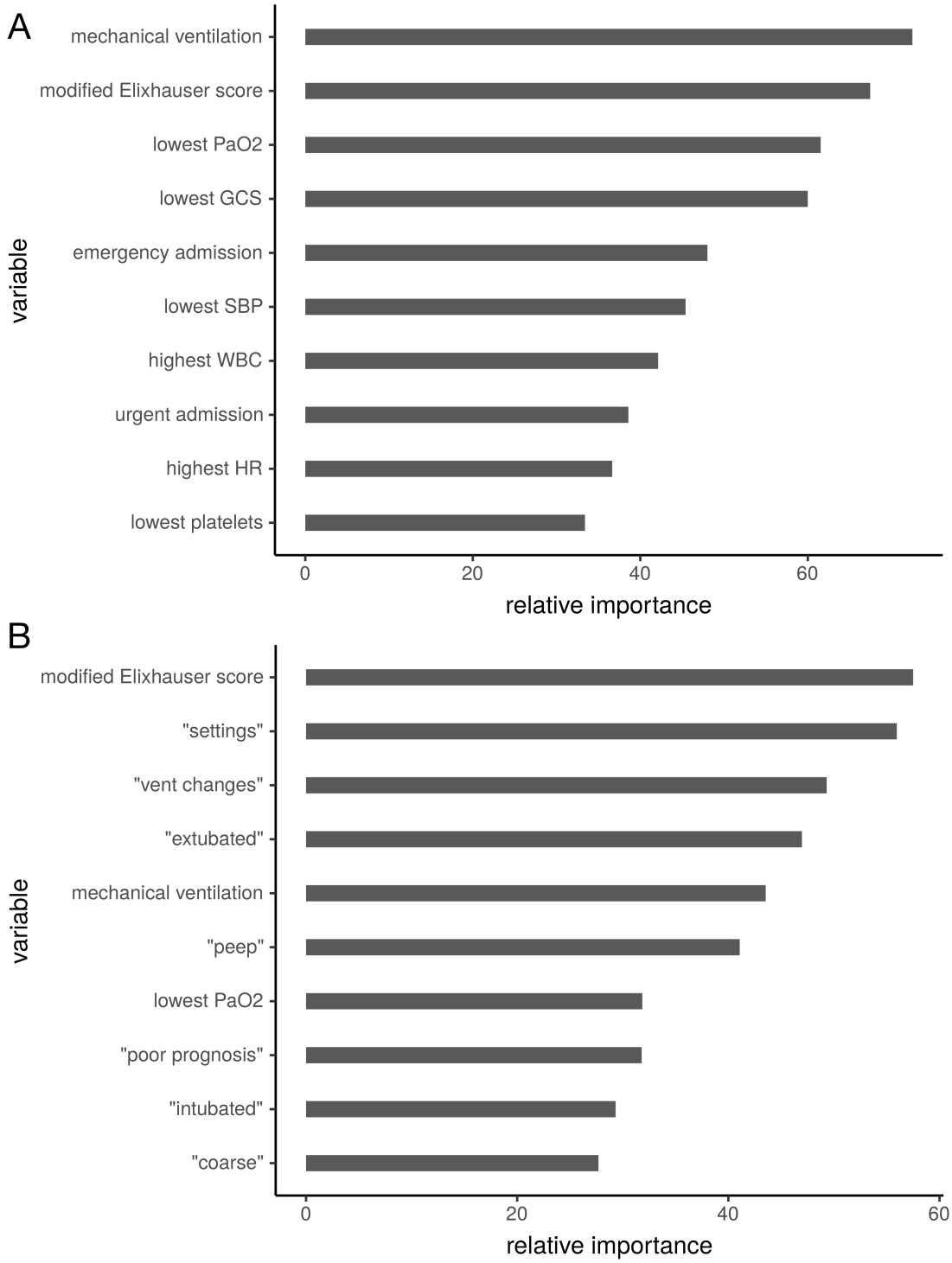
Supplemental figure e1

Supplemental figure e1: The sampling and cross-validation strategy used for model development and testing based on Friedman et al. (16)



Supplemental figure e2

Supplemental figure e2: The 10 most important variables by mean relative variable importance across all models using structured data only (A), and both structured and unstructured data (B). Variables representing terms derived from unstructured free-text data are presented within double quotation marks. Abbreviations: PaO2 = partial pressure of oxygen; GCS = Glasgow coma scale; SBP = systolic blood pressure; WBC = white blood cell count; HR = heart rate.


A
mechanical ventilation
modified Elixhauser score
lowest PaO2
lowest GCS
emergency admission
lowest SBP
highest WBC
urgent admission
highest HR
lowest platelets
variable
0
20
40
60
relative importance
B
modified Elixhauser score
"settings"
"vent changes"
"extubated"
mechanical ventilation
"peep"
lowest PaO2
"poor prognosis"
"intubated"
"coarse"
variable
0
20
40
60
relative importance

Secondary Analysis Results

Restricted 24-hour time horizon

	Unstructured data only		Structured and unstructured data	
Model Types	Training	Testing	Training	Testing
Logistic regression	0.80 (0.79 - 0.80)	0.80 (0.79 - 0.81)	0.88 (0.88 - 0.89)	0.84 (0.83 - 0.85)
Elastic net regression	0.80 (0.79 - 0.80)	0.80 (0.79 - 0.81)	0.87 (0.87 - 0.88)	0.86 (0.84 - 0.87)
Random forests	1.00 (1.00 - 1.00)	0.82 (0.80 - 0.83)	1.00 (1.00 - 1.00)	0.86 (0.85 - 0.87)
Gradient boosting machines	0.89 (0.88 - 0.89)	0.82 (0.81 - 0.84)	0.94 (0.94 - 0.95)	0.88 (0.87 - 0.89)

In-hospital death as primary outcome

	Unstructured data only		Structured and unstructured data	
Model Types	Training	Testing	Training	Testing
Logistic regression	0.83 (0.82 - 0.84)	0.83 (0.81 - 0.85)	0.93 (0.92 - 0.94)	0.83 (0.81 - 0.85)
Elastic net regression	0.83 (0.82 - 0.84)	0.83 (0.81 - 0.85)	0.92 (0.91 - 0.92)	0.86 (0.85 - 0.88)
Random forests	1.00 (1.00 - 1.00)	0.85 (0.83 - 0.86)	1.00 (1.00 - 1.00)	0.87 (0.85 - 0.88)
Gradient boosting machines	0.90 (0.90 - 0.91)	0.85 (0.84 - 0.87)	0.98 (0.98 - 0.98)	0.89 (0.88 - 0.90)

In-hospital death as primary outcome and restricted 24-hour time horizon

	Unstructured data only		Structured and unstructured data	
Model	Training	Testing	Training	Testing
Logistic regression	0.83 (0.82 - 0.83)	0.84 (0.82 - 0.85)	0.93 (0.92 - 0.93)	0.82 (0.80 - 0.84)
Elastic net regression	0.83 (0.82 - 0.84)	0.84 (0.82 - 0.85)	0.91 (0.91 - 0.92)	0.86 (0.84 - 0.87)
Random forests	1.00 (1.00 - 1.00)	0.84 (0.82 - 0.86)	1.00 (1.00 - 1.00)	0.87 (0.85 - 0.88)
Gradient boosting machines	0.92 (0.91 - 0.93)	0.85 (0.83 - 0.86)	0.97 (0.97 - 0.98)	0.89 (0.87 - 0.90)

In-hospital death and ICU length of stay $\geq$ 21 days

	Unstructured data only		Structured and unstructured data	
Model	Training	Testing	Training	Testing
Logistic regression	0.81 (0.80 - 0.82)	0.81 (0.80 - 0.83)	0.91 (0.90 - 0.92)	0.85 (0.83 - 0.86)
Elastic net regression	0.81 (0.80 - 0.82)	0.81 (0.80 - 0.83)	0.90 (0.89 - 0.91)	0.87 (0.86 - 0.88)
Random forests	1.00 (1.00 - 1.00)	0.83 (0.82 - 0.85)	1.00 (1.00 - 1.00)	0.87 (0.85 - 0.88)
Gradient boosting machines	0.92 (0.91 - 0.93)	0.84 (0.83 - 0.86)	0.97 (0.96 - 0.97)	0.89 (0.88 - 0.90)

Without using modified Elixhauser score for comorbidity adjustment

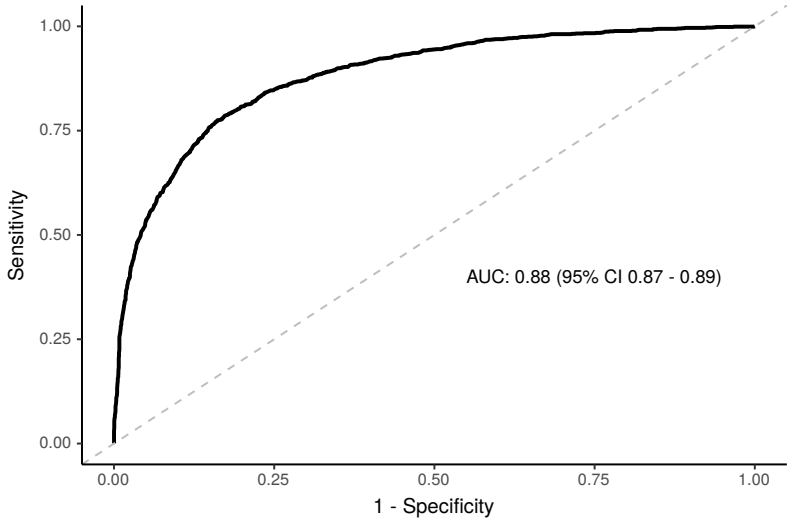
	Unstructured data only		Structured and unstructured data	
Model	Training	Testing	Training	Testing
Logistic regression	0.78 (0.77 - 0.79)	0.77 (0.76 - 0.79)	0.89 (0.88 - 0.90)	0.86 (0.84 - 0.87)
Elastic net regression	0.78 (0.77 - 0.79)	0.77 (0.76 - 0.79)	0.88 (0.87 - 0.89)	0.87 (0.85 - 0.88)
Random forests	1.00 (1.00 - 1.00)	0.81 (0.79 - 0.82)	1.00 (1.00 - 1.00)	0.87 (0.86 - 0.88)
Gradient boosting machines	0.88 (0.87 - 0.89)	0.81 (0.80 - 0.83)	0.95 (0.95 - 0.96)	0.89 (0.88 - 0.90)

Parsimonious model with 25 input variables only

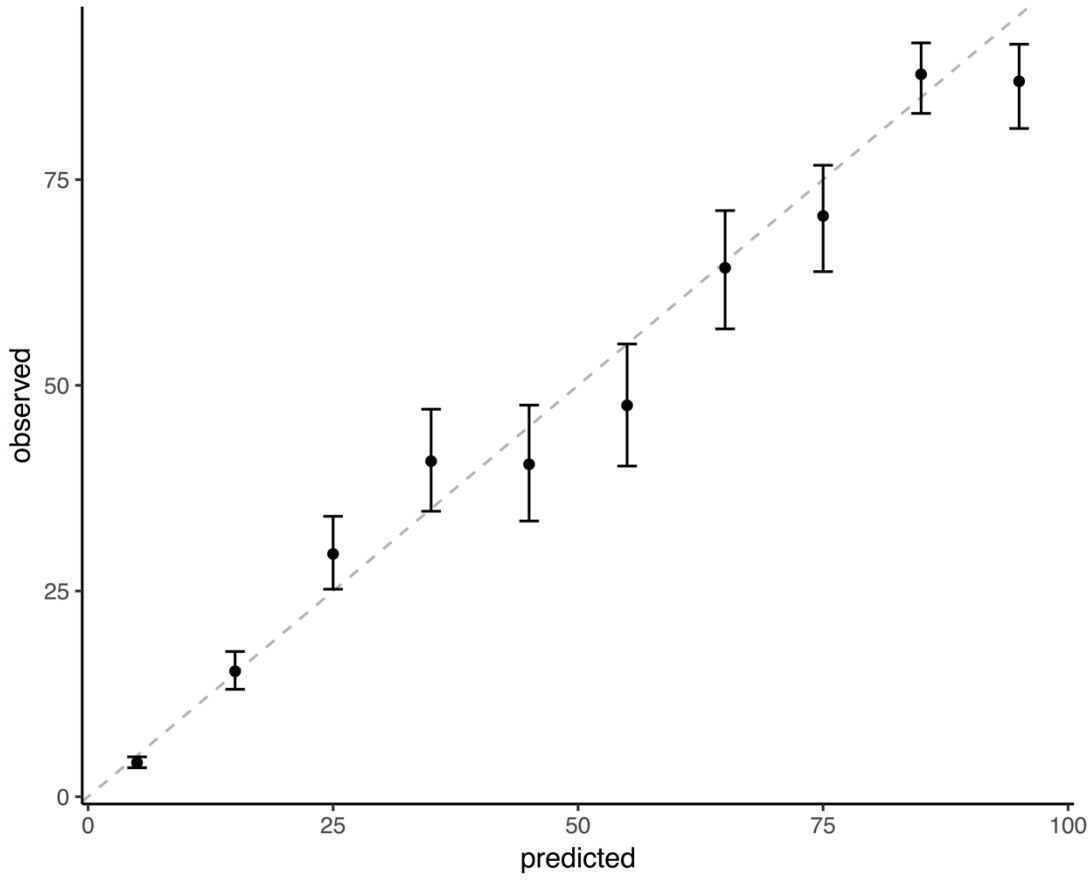
This sensitivity analysis used only the following variables:

Variable
“coarse”
“diet”
“extubate”
“extubated”
“extubation”
“intubated”
“oob”
“peep”
“settings”
“to floor”
“transfer”
“vent changes”
age
highest creatinine
highest heart rate
highest temperature
highest total bilirubin
highest WBC
lowest PaO2
lowest platelets
lowest systolic blood pressure
mechanical ventilation
modified Elixhauser score
urine output (cc/kg/hr)
word count

Supplemental figure e3: Receiver operating characteristic curve of a parsimonious gradient boosting machine model using only the 25 most predictive structured and unstructured variables from the equivalent model primary analysis.



Supplemental figure e4: Calibration plot of a parsimonious gradient boosting machine model using only the 25 most predictive structured and unstructured variables from the equivalent model primary analysis.



Supplemental table e1

Terms chosen *a priori* by the investigators to include as potential covariates in the prediction model with hypothesized likelihood of association with in-hospital mortality or long ICU stay.

Term
code status
comfort care
comfortable
death
die
doesn't want
family meeting
family
goals of care
high risk
hospice
mortality
mortal
pall care
palliation
palliative care
palliative
poor prognosis
survival
wishes
wouldn't want

Supplemental table e2

The 500 most predictive terms in the text based on the training sample only, sorted by the absolute value of the estimated log odds of their association (integer counts of occurrences within the text) with the composite outcome of in-hospital mortality or ICU length of stay ≥ 7 days.

Term	Log odds	Term	Log odds
wouldnt want	1.2349049	made	0.2092553
palliation	-1.2283002	extubation	-0.2037455
poor prognosis	1.0261102	copd	0.2018440
pall care	-1.0014613	tylenol	-0.2018124
hospice	0.7446242	ffp	0.2007944
goals of care	0.6002987	oob	-0.1967852
jaundice	0.4093547	pvc	0.1965917
palliative	0.3727337	tcurrent 37	-0.1945104
hcts	-0.3722134	weaned to	-0.1928535
liver	0.3670741	code	0.1928329
oob to	-0.3474088	rhythm o2 delivery	-0.1887645
comfort care	0.3418602	intubated	0.1875484
pupil	0.3169450	increasing	0.1864842
afib	0.3163799	teaching	-0.1847027
extubated	-0.2666501	ptt inr 14	-0.1823858
call out	-0.2657204	99	-0.1802441
coccyx	0.2559220	no seizure	-0.1775675
for comfort	0.2544545	coags	0.1762576
albumin	0.2529668	line placed	0.1737880
withdraws	0.2475301	us	0.1727993
meeting	0.2327209	confusion	0.1725332
nrb	0.2301489	78	0.1708512
wbc 10	-0.2286248	30cc	0.1691497
sit	-0.2267999	neuro exam	0.1687781
stimuli	0.2206194	higher	-0.1682006
levo	0.2183787	family meeting	0.1667954
to floor	-0.2171613	condition	0.1667366
started for	0.2108040	crackles	0.1661240
diet	-0.2101539	vomited	0.1660709
obese	-0.2093904	resp	0.1659120
		vit	0.1638364

and draining	0.1622617	urine culture	-0.1196400
tachypneic	0.1595002	awoke	-0.1188748
resp rate	0.1589393	creat	0.1181299
wean as	-0.1560710	colored	0.1178844
atelectasis	-0.1547596	qid	0.1168911
discuss	0.1534775	compression	0.1162134
hosp	0.1532891	name5	-0.1145222
transfer to floor	-0.1523745	amber urine	0.1133570
status	0.1505920	firm	0.1131818
deep breathing	-0.1502667	sediment	-0.1130109
clear yellow	-0.1496122	lr	-0.1127376
coarse	0.1490516	the key portions	-0.1119715
resection	0.1467760	ongoing	-0.1110143
decision	0.1441715	ngt	0.1106843
inch admission weight	-0.1433891	sedated on propofol	-0.1102541
abp	0.1431900	they	0.1077453
abd is soft	-0.1407238	response to	0.1068395
sent for	0.1393743	cool	0.1067955
son	0.1391633	well	-0.1062959
poor	0.1380455	adequate	-0.1061267
to self	0.1370443	advanced	-0.1060617
csm	-0.1357409	to sleep	-0.1057657
sat 97	-0.1356393	coumadin	-0.1056472
clamped	0.1350838	bs coarse	-0.1042164
breakdown	0.1348516	pt not	0.1041013
endo insulin	-0.1319175	lung	0.1025791
encouraged	-0.1298372	air	-0.1015915
painful	0.1285621	bm this	-0.0999619
doppler	0.1281939	death	0.0993629
on room air	-0.1235033	blood tinged	0.0985198
pain	-0.1231860	called out to	-0.0972372
will hold	0.1229751	close	-0.0970817
and increased	0.1222616	monitoring	-0.0969307
lift	-0.1213521	incisional pain	-0.0969305
male with	0.1210935	decreased	0.0967698
integ	0.1201746	amiodarone	0.0964373

none spo2	-0.0962166	brisk	-0.0811818
hemorrhage	0.0961836	his wife	0.0809158
ew	-0.0956478	clear yellow urine	-0.0809050
hemodynamically stable	-0.0947767	to keep sbp	0.0806568
better	-0.0944681	nsr	-0.0806411
rr 12	-0.0937882	40cc hr	0.0803064
wake	-0.0937284	tolerating	-0.0802265
stable	-0.0919050	on 40	-0.0788982
jp	-0.0916081	upper	0.0786752
brown	0.0913397	fever	-0.0785291
hr and	-0.0910317	await	0.0775587
able to	-0.0909471	remains intubated and	0.0775256
depression	-0.0907620	section	-0.0775219
noted to	0.0904827	amounts of thick	-0.0770698
absent left	-0.0903952	ground	0.0770367
bil	-0.0895086	children	0.0767200
syndrome	0.0893703	monitor vs	-0.0766305
wnl	-0.0892768	resolved	-0.0760909
able to wean	-0.0882943	dl ldh	-0.0757747
sb	-0.0881672	open	0.0755961
with family	0.0877257	throughout	0.0750056
250cc	0.0876955	low grade	-0.0748376
iv access	-0.0876190	po	-0.0739339
answer	-0.0875148	sob and	0.0732318
very pleasant	-0.0873701	sounds present no	-0.0728116
became	0.0873367	until	-0.0727457
of clear	-0.0862285	chest ct	0.0723189
unresponsive	0.0862083	completed	-0.0721254
wires	-0.0855609	please see carevue	0.0720350
form	-0.0846658	covered with	0.0714437
done	0.0843260	for	0.0713078
155	0.0839316	wean vent	-0.0708337
ace	-0.0829121	ue	0.0707617
received total	-0.0822732	would be	0.0707598
effusion	0.0819425	respiratory care pt	-0.0707289
transfer	-0.0812290	to make	0.0705645

labile	0.0702871	appropriately	-0.0617282
q2	-0.0702210	adequate amts	-0.0614802
femoral	0.0690776	wants	0.0614676
no changes	0.0686173	trauma	-0.0613460
cooperative	-0.0685695	central	0.0610166
clear	-0.0683661	and was	-0.0608678
97 abg	-0.0683448	70 90	-0.0607148
awake	-0.0678976	min spo2 99	-0.0605990
this am with	-0.0678888	extubate	-0.0603336
12 am	-0.0678574	dopamine	0.0602402
to decrease	0.0678325	placement	0.0590904
unchanged	0.0674402	11	0.0588919
hcp	0.0672108	marginal	0.0581989
2u	0.0671761	no sob	-0.0581938
bowel regimen	-0.0666411	generalized	0.0579898
dilantin	0.0665557	too	-0.0579591
soft nontender	-0.0665553	lower	0.0579456
wbc hct	-0.0665355	on cmv	0.0579200
oriented x3	-0.0662822	worsening	0.0577038
to painful stimuli	0.0658659	left	0.0571774
note neuro	0.0658275	repleted	-0.0571342
pearl	-0.0655045	as tolerated	-0.0567889
most recent	0.0653622	bm	0.0567600
benign	-0.0653472	ss	-0.0565686
pupils equal and	-0.0647768	weaned and	-0.0558482
uf heparin stress	-0.0645102	hydration	0.0555366
father	-0.0644494	rising	0.0551725
ci	-0.0639538	bronch	0.0548900
and on	-0.0630917	and versed	-0.0548090
boluses	0.0630118	due	0.0543968
scan	0.0629298	placed on	0.0542288
strong	-0.0628221	today	0.0539958
mostly	0.0627935	associated	-0.0539259
and fentanyl	-0.0625297	healthy	-0.0532552
services	-0.0623281	fluids	-0.0526945
to 70	0.0622012	by name	-0.0523174

sleeping	-0.0522945	plans	0.0449128
state	0.0522756	plts	0.0447652
colace	-0.0521106	revealed	0.0447195
weaning	-0.0520980	hr 80	-0.0446010
pull	-0.0519167	son name	0.0444710
to 30	0.0518003	pt more	-0.0444255
rr 20	0.0513603	on ra	-0.0442920
bp and	0.0512538	maintain map	0.0441844
of ns	0.0511755	occas	0.0441608
pip	0.0506808	improved	-0.0440403
uti	-0.0503546	grade temp	-0.0430391
tolerate	0.0498602	to increase	0.0429691
rr 22	0.0496428	stable on	0.0429554
up in	-0.0494854	as tol	-0.0420957
drop	0.0491632	sats 93	0.0420402
increased	0.0487968	at bedside	0.0418982
voiding	-0.0482559	psych	-0.0418660
easily	-0.0480021	is for	-0.0417775
reddened	0.0478586	foley	0.0413868
foley cath	-0.0477538	hr to	-0.0412125
final	0.0476737	mother	-0.0411882
nods	-0.0471518	events pt	-0.0405802
bp 120	-0.0469787	tcurrent 36 98	-0.0404490
and her	-0.0468680	spec	0.0400962
due to	0.0467913	needed for	-0.0398785
hypoxia	0.0467648	50 mcg	-0.0398332
equal and	-0.0467609	sluggish	0.0397646
normal	-0.0467017	states that	-0.0397444
the ed	-0.0466907	distended	0.0395758
bolus and	0.0461039	strength	-0.0395573
follows	-0.0458142	gtt off	-0.0388891
titrate	0.0454427	pacer	-0.0387622
draining	0.0454193	pt states	-0.0380118
pulses	-0.0453485	cefazolin	-0.0378050
increasingly	0.0449944	rr 17	-0.0375026
total	-0.0449747	within	-0.0374755

spo2 99	-0.0373984	for hypotension	0.0318674
done and	-0.0373487	suctioned small	0.0314345
phone	-0.0373230	possibly	0.0311221
closely	0.0372033	cleared	-0.0306002
be	0.0371586	pantoprazole	-0.0303851
chest	-0.0370931	free	-0.0297468
rue	0.0369502	rr	0.0297097
fall	0.0366570	palpable pulses	-0.0294901
social work	-0.0366264	in bed	-0.0294843
as ordered	-0.0366235	to osh	-0.0292791
spo2 100	-0.0365425	vss	-0.0291775
bilat	0.0364992	spent 35	-0.0288156
spoke with	0.0359904	will follow	-0.0286017
for mod	-0.0359482	admission	-0.0285645
peep	0.0354804	ct scan	0.0284059
03	-0.0353728	mn	-0.0283857
beta	-0.0352711	maps	0.0282046
60 70	-0.0349503	switched	-0.0281033
with normal	-0.0347339	wishes	0.0279985
this evening	0.0346727	and the	0.0278162
ischemic	0.0343352	resp lungs	-0.0277797
name stitle	0.0343096	reports	-0.0272953
98 hr	-0.0342200	sinus rhythm	-0.0270242
head ct	0.0340702	guiac	0.0265847
and to	0.0339392	daughter	0.0265454
on 4l nc	0.0336610	tomorrow	-0.0263670
alert	-0.0334092	spo2 98	-0.0263389
improvement	0.0333021	on lopressor	-0.0262139
family in	0.0331802	gtt started	0.0260683
support and	-0.0327311	and his	-0.0251238
12 00	-0.0323517	tmax 100	-0.0250561
20 meq	-0.0323008	resume	-0.0248753
sedation and	-0.0321865	minimally	0.0248484
address	0.0321749	at all	0.0245230
abdomen	0.0320145	from or	-0.0244264
weaned	-0.0319301	70s	0.0242131

and daughter	0.0241661	uo	0.0191854
db	-0.0238271	tele	-0.0191684
additional	-0.0236754	vanco	0.0190041
ice	-0.0236097	ed	-0.0189523
obtain	0.0232972	aware	0.0184764
nose	-0.0230885	unlabored	-0.0181546
encourage	-0.0229505	abd pain	0.0181064
in good	-0.0225792	cont to monitor	-0.0176443
ciwa	-0.0225748	discharge	-0.0174784
ra	-0.0224272	barrier	-0.0174464
weakness	0.0224097	to 100	-0.0173078
toilet	-0.0223856	room air	-0.0170826
overbreathing	0.0222335	with increased	-0.0168565
hrs	-0.0219154	30cc hr	0.0167193
for meds	0.0219052	bronchial	0.0165429
be done	0.0218213	placed	0.0162557
despite	0.0217501	occ	0.0162517
bloody	0.0216054	thick white	-0.0162263
hr of	-0.0214436	map 60	0.0157581
when she	-0.0213723	reported	-0.0157530
elevated	0.0212575	code full	-0.0156433
pt npo	-0.0212376	if	-0.0151594
units of	-0.0211688	vent changes	0.0150982
per orders	-0.0210212	endo insulin gtt	-0.0150445
sr	-0.0209813	loss	0.0150173
attempted	0.0207807	asleep	-0.0149252
it was	0.0207617	3l	0.0148696
would like	0.0205967	freq	0.0148670
emotional	0.0202373	na	0.0147766
no signs	-0.0202150	for the	-0.0145904
in micu	0.0201256	sounds diminished	0.0145672
neuro pt alert	-0.0197525	transported	0.0142698
clear bilaterally	-0.0196722	mental	0.0142347
hours	-0.0195959	settings	0.0140384
questions	-0.0193225	working	-0.0136843
insulin gtt	-0.0193174	bp	0.0134556

or	-0.0132712	hyperglycemia	-0.0128553
with good	-0.0131280	history of	-0.0127285
discussion	0.0130703	67	-0.0126486
and wean	-0.0129695		

Supplemental table e3

Characteristic, n (%)	All	Died or ICU LOS ≥ 7 days	
		Yes	No
Modified Elixhauser score, median (IQR)	8 (2 – 15)	12 (6 – 19)	6 (1 – 13)
Clinical deterioration			
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)
WBCs, 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)
PaO2, 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)
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 (mmHg), lowest	89 (79 – 101)	83 (70 – 94)	91 (81 – 102)
Heart rate (beats per minute), 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			
Total raw word count, median (IQR)	1,023 (562 – 2,266)	1,410 (956 – 2,292)	899 (501 – 2,247)
Number of notes, median (IQR)	6 (4 – 11)	8 (6 – 12)	5 (3 – 10)

Supplemental table e4

Variable	Missingness, n (%)	Imputed mean
Bilirubin	11,444 (44.1)	1.54
Creatinine	40 (0.2)	1.66
Glasgow coma scale	6,964 (26.8)	9.61
Heart rate (bpm)	418 (1.6)	69.1
Potassium	115 (0.4)	4.68
PaO2	1,4021 (54.0)	103.0
Platelets	57 (0.2)	190.3
Systolic blood pressure (mmHg)	418 (1.6)	86.5
Temperature (C)	515 (2.0)	35.5
Urine output (cc/kg/hr)	8,826 (34.0)	0.85
White blood cells	84 (0.3)	14.4
Weight (kg)	5,422 (21.0)	83.3

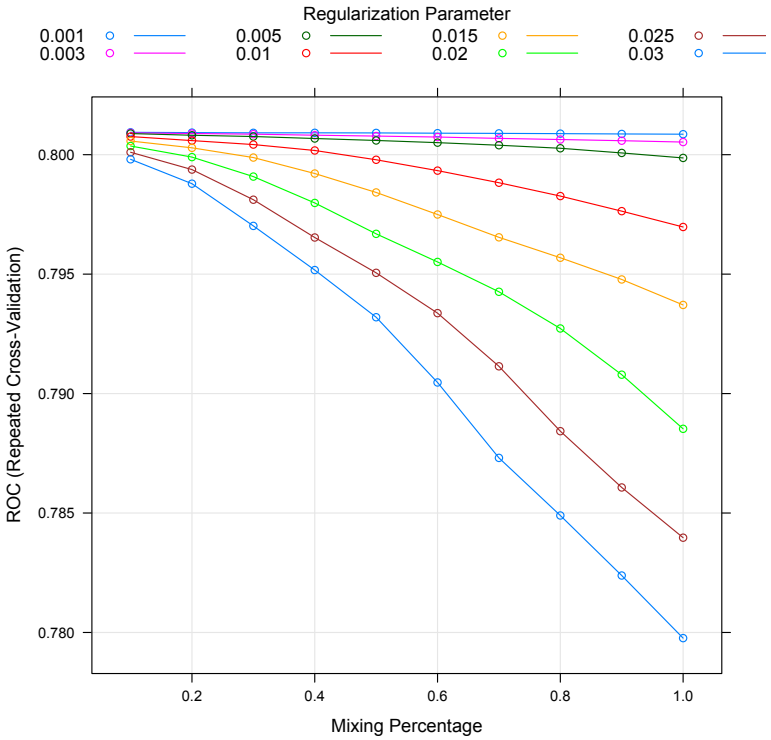
Supplemental table e5

Machine learning model types and tuning parameters.

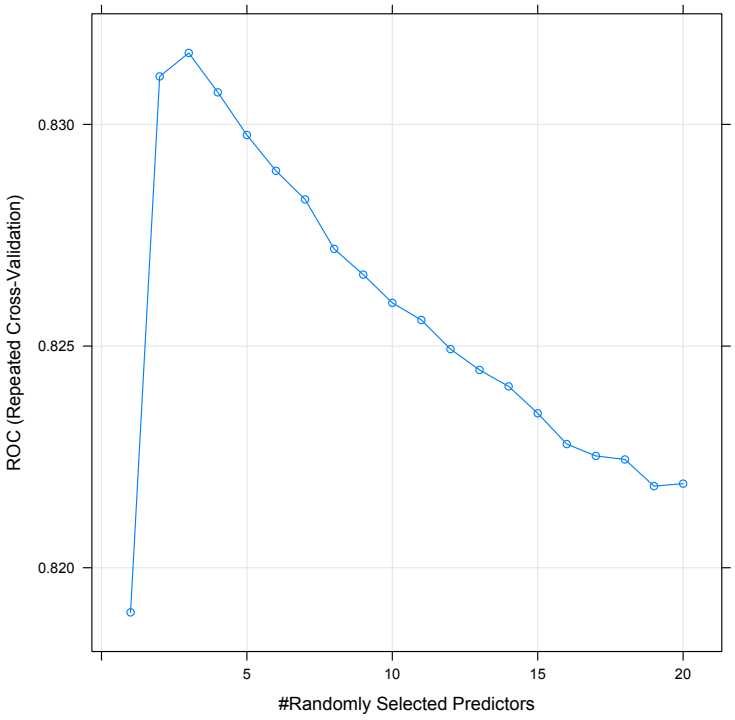
Model family	Model type	Tuning parameters	Software package
Logistic regression	Traditional	None	Base R
	Elastic net	Mixing, regularization	glmnet
Tree-based	Random forest	Variables per node	randomForest
	Gradient boosting machine	Number of trees, number of splits, regularization, minimum observations per node	gbm

Cross-Validation Performance in the Primary Analysis

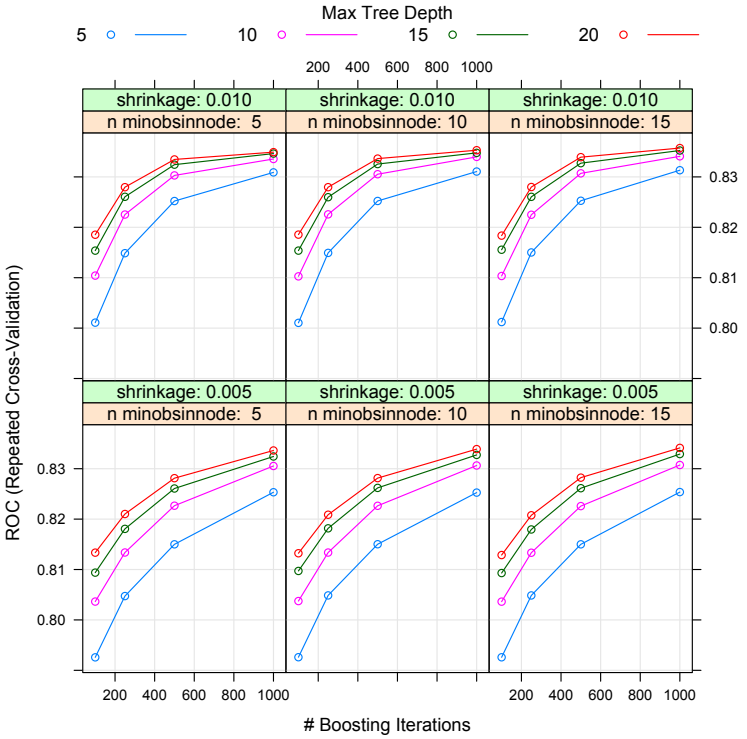
Supplemental figure e5: Elastic Net model using structured data only



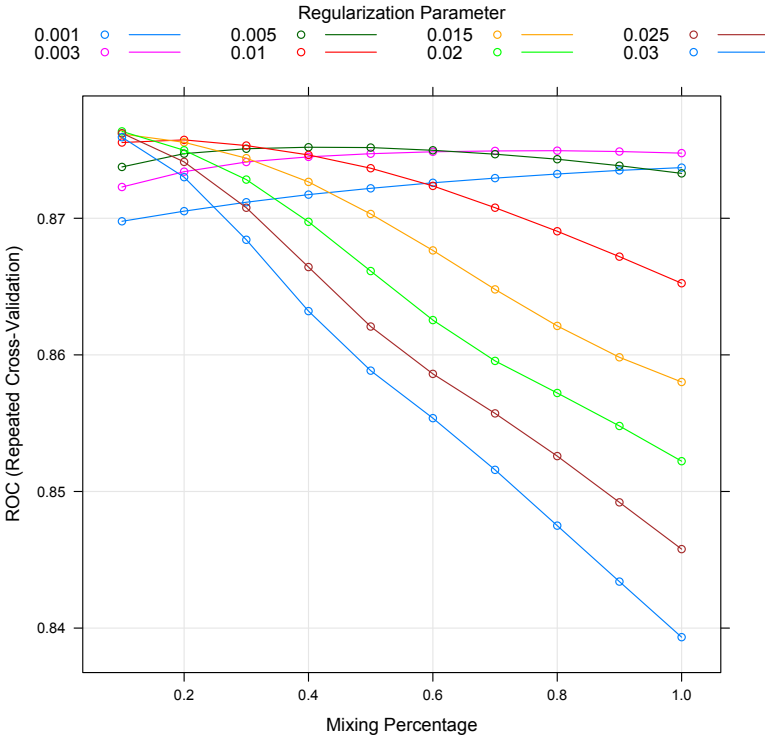
Supplemental figure e6: Random Forest model using structured data only



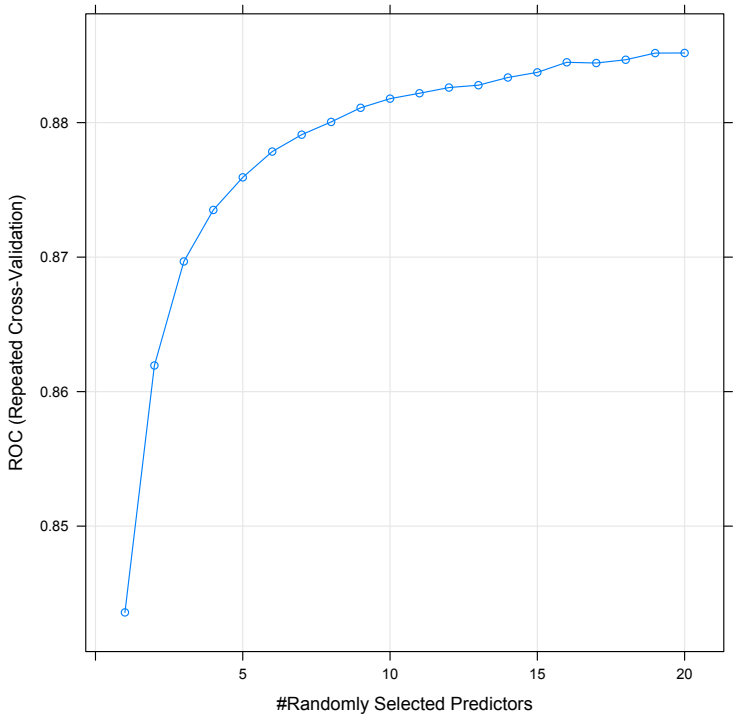
Supplemental figure e7: Gradient Boosting Machine model using structured data only



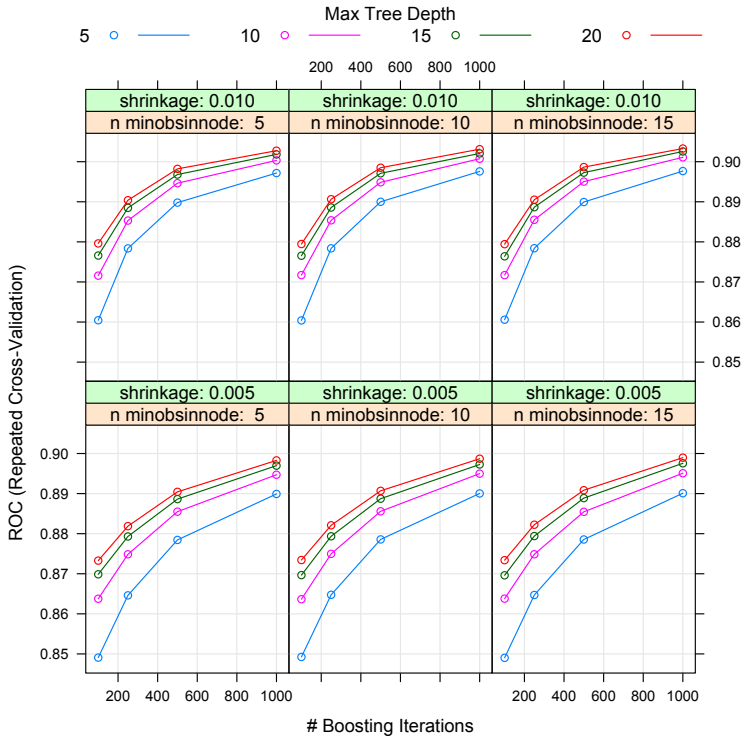
Supplemental figure e8: Elastic Net model using structured and unstructured data



Supplemental figure e9: Random Forest model using structured and unstructured data

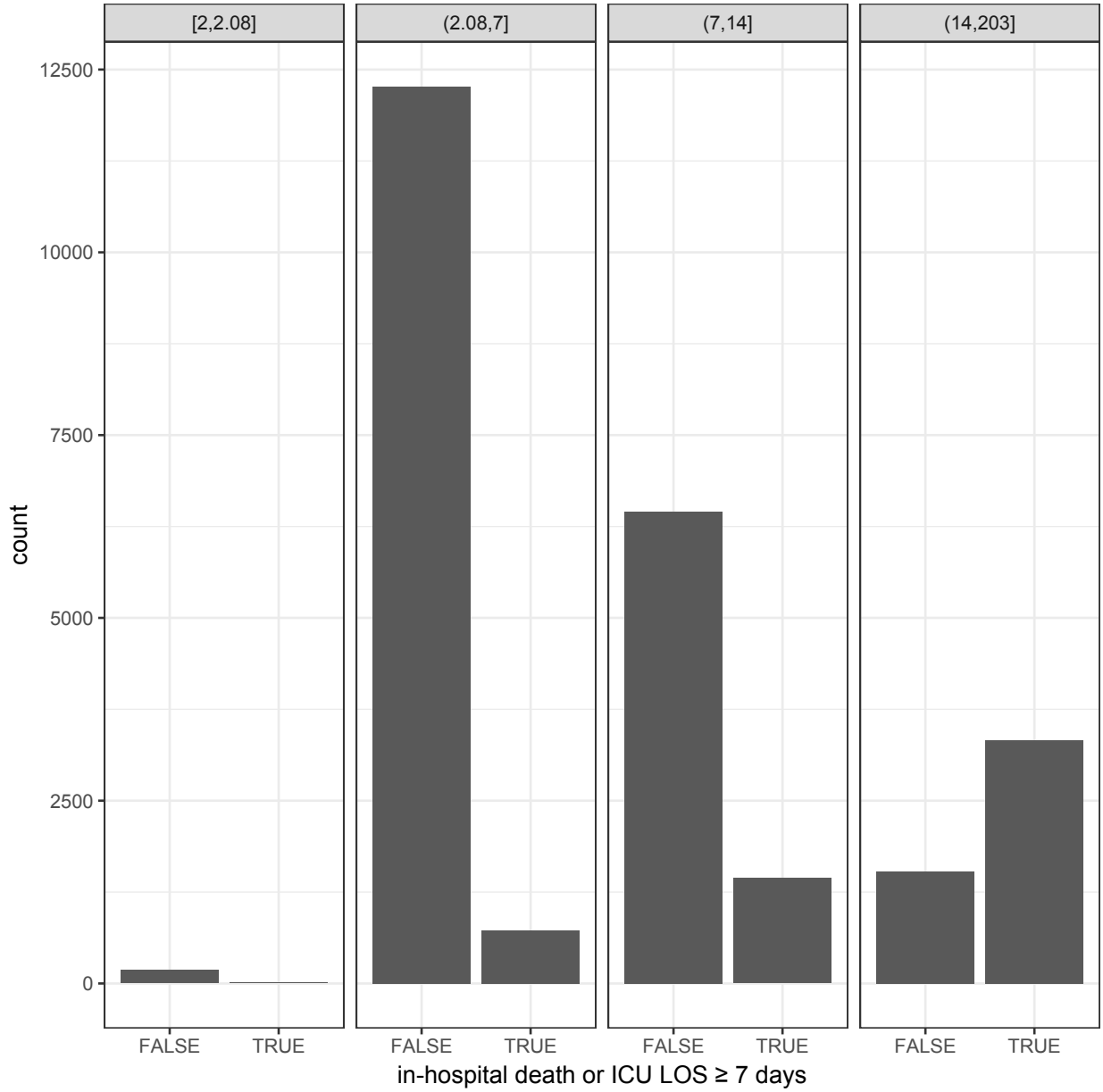


Supplemental figure e10: Gradient Boosting Machine model using structured and unstructured data



Model Performance by Hospital Length of Stay

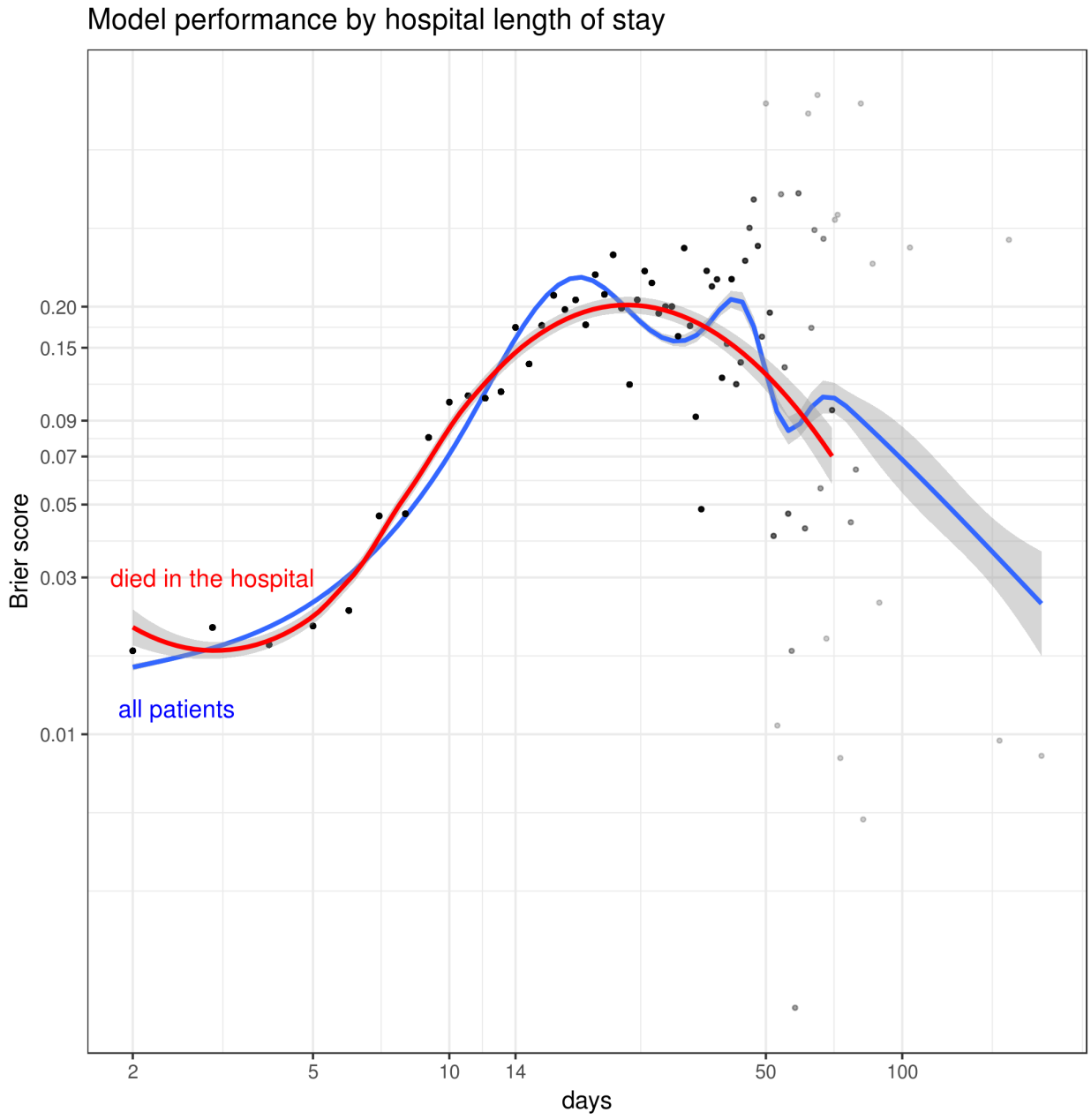
Supplemental figure e11: Mortality and ICU length of stay by hospital length of stay (days)



Supplemental table e6: Performance of gradient boosting machine model with text data in the testing sample by hospital length of stay

Hospital length of stay (days)	C-statistic	95% confidence interval	Observations (n)
[2, 2.08]	0.953	0.937 – 0.968	53
(2.08, 7]	0.830	0.805 – 0.854	3,268
(7, 14]	0.793	0.767 – 0.819	1,959
(14, 203]	0.929	0.785 – 1.000	1,206

Supplemental figure e12: LOESS plot with 95% confidence intervals of performance of the gradient boosting machine model using structured and unstructured data as measured by the Brier score over the total hospital length of stay. Both the x- and y-axes have been transformed to a log_{10} scale.



Additional Model Performance Characteristics

Supplemental table e7: Sensitivity (Sens), Specificity (Spec), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and F1 Score for all logistic regression (LR), elastic net regression (EN), random forest (RF), and gradient boosting machine (GBM) models using the held-out testing sample. Continuous predicted probabilities are considered positive for p >= 0.5.

	Unstructured data only					Structured and unstructured data				
Model	Sens	Spec	PPV	NPV	F1	Sens	Spec	PPV	NPV	F1
LR	0.954	0.301	0.840	0.628	0.894	0.946	0.539	0.888	0.720	0.916
EN	0.954	0.298	0.840	0.629	0.893	0.961	0.461	0.873	0.752	0.915
RF	0.956	0.339	0.847	0.667	0.899	0.962	0.468	0.875	0.760	0.916
GBM	0.944	0.401	0.859	0.650	0.899	0.950	0.558	0.892	0.745	0.921

Supplemental table e8: Sensitivity (Sens), Specificity (Spec), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and F1 Score for all logistic regression (LR), elastic net regression (EN), random forest (RF), and gradient boosting machine (GBM) models using the held-out testing sample. Continuous predicted probabilities are considered positive based on a probability threshold (Thresh) in each case determined by the Youden method.

	Unstructured data only						Structured and unstructured data					
Model	Thresh	Sens	Spec	PPV	NPV	F1	Thresh	Sens	Spec	PPV	NPV	F1
LR	0.184	0.766	0.685	0.386	0.919	0.513	0.184	0.789	0.805	0.512	0.936	0.621
EN	0.183	0.769	0.680	0.384	0.919	0.512	0.183	0.813	0.786	0.496	0.942	0.616
RF	0.229	0.751	0.737	0.425	0.920	0.543	0.269	0.759	0.850	0.567	0.931	0.649
GBM	0.205	0.728	0.770	0.450	0.916	0.556	0.213	0.774	0.854	0.579	0.936	0.662

References

1. Simon N, Friedman J, Hastie T, et al.: Regularization paths for Cox's proportional hazards model via coordinate descent. [Internet]. *Journal of statistical software* 2011; 39:1–13Available from: http://www.ncbi.nlm.nih.gov/pubmed/27065756

2. Liaw A, Wiener M: Classification and Regression by randomForest [Internet]. *R News* {2002}; 2:18–22Available from: {http://CRAN.R-project.org/doc/Rnews/}

3. Ridgeway G, contributions from others: gbm: Generalized Boosted Regression Models [Internet]. {2015}. Available from: {https://CRAN.R-project.org/package=gbm}

4. McCormick P, Joseph T: medicalrisk: Medical Risk and Comorbidity Tools for ICD-9-CM Data [Internet]. {2016}. Available from: {https://CRAN.R-project.org/package=medicalrisk}

5. Walraven C van, Austin PC, Jennings A, et al.: A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. [Internet]. *Medical care* 2009; 47:626–633Available from: http://www.ncbi.nlm.nih.gov/pubmed/19433995

6. Elixhauser A, Steiner C, Harris DR, et al.: Comorbidity measures for use with administrative data. [Internet]. *Medical care* 1998; 36:8–27Available from: http://www.ncbi.nlm.nih.gov/pubmed/9431328

7. Quan H, Sundararajan V, Halfon P, et al.: Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data. [Internet]. *Medical care* 2005; 43:1130–1139Available from: http://www.ncbi.nlm.nih.gov/pubmed/16224307

8. Moreno RP, Metnitz PGH, Almeida E, et al.: 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. [Internet]. *Intensive care medicine* 2005; 31:1345–1355Available from: http://www.ncbi.nlm.nih.gov/pubmed/16132892

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 [Internet]. *Critical care medicine* 2006; 34:1297–1310Available from: http://www.ncbi.nlm.nih.gov/pubmed/16540951

10. Matt Dowle and Arun Srinivasan: data.table: Extension of 'data.frame' [Internet]. {2017}. Available from: {https://CRAN.R-project.org/package=data.table}

11. Hadley Wickham: ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; {2009}.

12. Kuhn M, Other contributors.: caret: Classification and Regression Training [Internet]. {2016}. Available from: {https://CRAN.R-project.org/package=caret}

13. R Core Team: R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.R-project.org/

14. Bird S, Klein E, Loper E: Natural language processing with python: Analyzing text with the natural language toolkit. " O'Reilly Media, Inc." 2009.

15. Pedregosa F, Varoquaux G, Gramfort A, et al.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011; 12:2825–2830

16. Friedman J, Hastie T, Tibshirani R: The elements of statistical learning. Springer series in statistics Springer, Berlin; 2001.

17. 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. In: A22. health services highlights in critical care. Am Thoracic Soc; 2017. p. A1084–A1084.

18. Nadkarni PM, Ohno-Machado L, Chapman WW: Natural language processing: an introduction. *J Am Med Inform Assoc* 2011; 18:544-551.

19. Xu H, Stetson PD, Friedman C: A Study of Abbreviations in Clinical Notes. *AMIA Annual Symposium Proceedings*. 2007;2007:821-825.

20. Wang MD, Khanna R, Najafi N: Characterizing the Source of Text in Electronic Health Record Progress Notes. *JAMA Intern Med.* 2017;177(8):1212–1213.

21. Nikfarjam A, Emadzadeh E, Gonzalez G: Towards generating a patient's timeline: Extracting temporal relationships from clinical notes. J Biomed 2013;46:S40-S47.

22. 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.