

## Supplementary Online Content

Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 natural language processing methods for identification of bleeding among critically ill patients. *JAMA Netw Open*. 2018;1(6):e183451. doi:10.1001/jamanetworkopen.2018.3451

**eTable 1.** Features Selection for Machine Learning: n-grams Associated and Not Associated With Clinically Relevant Bleeding

**eTable 2.** Target Terms Used in Natural Language Processing Task to Identify Bleeding From Clinical Notes

**eTable 3.** Modifier Terms Used in Natural Language Processing Task to Identify Bleeding From Clinical Notes

**eTable 4.** Average Computational Time for Each Natural Language Processing Approach

**eTable 5.** “Bleeding Present” Phrases Identified by the Rules-Based Algorithm, but Not Identified by the Physician Reviewer

This supplementary material has been provided by the authors to give readers additional information about their work.

**eTable 1. Features selection for machine learning: n-grams associated and not associated with clinically relevant bleeding<sup>a</sup>**

Features associated with “bleeding absent”	Features associated with “bleeding present”
labetolol mg syringe mmhg cardiac output hypertrophic change make medications start spironolactone history significant diabetes significant diabetes pna patient drainage incision day omeprazole initial chest position seat room follow sputum culture grow small bilateral spiriva inh chest tube placement blood base intuba lastname continue	bleed gi gi bleed egd gib hematoma gastrointestinal bleed hemorrhage subdural bleed gib gi bleed gib ffp bleed gi bleed gastrointestinal bleed gi bleed gi units units units octreotide prbcs hematuria units units
<sup>a</sup> Feature selection was performed using chi-squared tests. The chi-square value was zero for all features associated with “bleeding absent.” Chi-square values range from 1.626 to 9.062 for features associated with “bleeding present”	

**eTable 2. Target terms used in natural language processing task to identify bleeding from clinical notes.**

Clinical Concept	Regular Expression
bleed	(?<!non)(?<!non )(?<!non-)(?<!re)(bleed(?!ing) bleeding(?!\stime))
blood loss	blood loss
blood per rectum	blood per rectum
bloody	(?<!non-)(?<!non)(?<!non )bloody
brbpr	brbpr
coffee grounds	coffee[- ](ground grounds)
ecchymosis	ecchymos[ie]s
epistaxis	epistaxis
exsanguination	exsanguination
gib	\b!gib\b
guaiac	((\bg gua?iac)([-])\s+)((pos(itive)?)\ +)\ guaiac\(\+\))
hematemesis	hematem[a-z]+
hematochezia	hematochezia
hematoma	hematoma
hematuria	hematuria
hemoperitoneum	hemoperitoneum
hemoptysis	hemoptysis
hemorrhage	(?<!splinter\s)hem{1,2}or{1,2}h{1,2}age?
ich	\bich
melena	mel[ae]n(alic)
ng lavage post	(ng ngt)\s+lavage\s+((positive) \ )\ +
ng lavage pre	((positive) \ )\ + (ng ngt) lavage
occult blood	(fecal\s+occult(\s+blood)?\ bob \bfob)\s+pos(itive)?
sah	sah
sdh	sdh
stool	(maroon red)\s+(stool bowel\s+movement bm)
vomitting	vomit[a-z]* blood

**eTable 3. Modifier terms used in natural language processing task to identify bleeding from clinical notes. A minus-sign (-) in the Rule column indicates that the modifier marked its associated target term as bleeding-absent. A plus-sign (+) indicates that the modifier confirmed its target term as bleeding-present.**

Clinical Concept	Regular Expression	Rule
ago	\bago\b	-
can result in	\bcan\s+result\s+in\b	-
cc	(cc: chief complaint:)	+
ddx	ddx	-
deny	denies denied denying	-
did not show	did not (show reveal)	-
episode	episode of	+
episodes	episodes of	+
found to have	found to have	+
here with	here (with w\\ w/)	+
history	((h/o) (h\\o) (hx ?of) history \bhx\b pmh)	-
history of	history\s+of	-
if	\bif\b	-
in the past	\bin\s+the\s+past\b	-
monitor	monitor( for)?	-
negative	\bnegative\b	-
negative dash	\(-)	-
never	\bnever\b	-
no	\bno\b	-
no evidence	no (evidence e\\o e/o)	-
no history	no ((h/o) (h\\o)) hx)	-
non	non(\s -)	-
not	not\b	-
now with	now (with w\\ w/)	+
possible	possible	-
presenting	presenting* (with w\\ w/)	+
presents	present[s e][d]* (with w\\ w/)	+
previous	previous	-
prior	\bprior\b	-
risk	risk of	-
rule out	(rule out rVo)	-
recent admission	recent\s+admission	-
suspicion	suspicion	-
transfuse	transfuse	-
unlikely	unlikely	-
unrevealing	unrevealing\s+for	-
versus	versus vs	-
watch	watch for	-
without	((without) (w\\o) (w/o))(?!scontrast)	-

**eTable 4. Average computational time<sup>a</sup> for each natural language processing approach.**

<b>Model</b>	<b>Notes per Second</b>
SVM	42.72
ET	5818.42
CNN	122.16
RB	3.09

<sup>a</sup>Computational time if calculated as the number of notes each model is able to process per second, as an average over the 660-note test set. These tests were run on a machine with a 4-core intel i7 CPU and 16 gigabytes of RAM.

SVM=support vector machine; ET=extra trees classifier; CNN=convolutional neural network; RB=rule-based model

**eTable 5. “Bleeding present” phrases identified by the rules-based algorithm, but not identified by the physician reviewer**

left upper extremity hematoma
with signs/symptoms of bleeding
loose melena
L second toe with hemtoma present
likely hemorrhage related to
bleed noted at cath site
coffee-ground emesis
guaiac + stools
Small hematoma
re-op for chest wall hematoma
diffuse ecchymoses, INR > 22.8
guaic + stools
persistent area of bleeding
bloody exudate
R groin with ecchymosis
Anemia, acute, secondary to blood loss (Hemorrhage, Bleeding)
stools were guaiac positive
excess amounts of postoperative bleeding
left parietal subgaleal hematoma present
line site bleeding
subdural hematoma on L
consistent with alveolar hemorrhage