Supplementary Table 2. *NLP System Description*

Author(s), Year	NLP System Used	Type of NLP Method Used	Standard Terminology Used	Number of Documents used	Reported NLP System Performance
Bjarnadottir et al, 2018[53]	MySQL	Rule-based	SNOMED, ICD-9, NANDA, LOINC	1,046,053	Not reported
Bjarnadottir et al, 2019[52]	Intellij for Java; AutoMap	Rule-based	SNOMED, ICD-9, LOINC	862,715	Not reported
Boyd et al, 2018[9]	MedLEE	Hybrid (Bayesian)	UMLS, NANDA, NIC, NOC	40,719	Not reported
Conway et al, 2019[51]	Moonstone	Rule-based	SNOMED, UMLS	52,304	The system achieved positive predictive value (i.e., precision) scores ranging from 0.66 (homeless/marginally housed) to 0.98 (lives at home/not homeless), accuracy scores ranging from 0.63 (lives in facility) to 0.95 (lives alone), and sensitivity (i.e., recall) scores ranging from 0.75 (lives in facility) to 0.97 (lives alone).
De Silva et al, 2021[17]	LASSO	Rule-based	Not reported	408,560	Not applicable
Fralick et al, 2021[48]	Not reported	Rule-based	Keyword search	Not reported	Not reported
Galatzan et al, 2021[47]	LIWC	Rule-based	Not reported	Not reported	Not reported
Gundlapalli et al, 2017[50]	V3NLP	Rule-based	Not reported	1,595	The overall recall was 75% and positive predictive value was 99% on the

					training set; on the testing set, the recall was 72% and positive predictive value was 98%. The performance on extracting urinary symptoms (including fever) was high with recall and precision greater than 90%.
Hajihashemi et al, 2013[59]	Metamap	Rule-based	UMLS	626	Not applicable
Harkanen et al, July 2019[46]	SAS Text miner & SAS Enterprise Miner 13.2	Hybrid	Not reported	1,012	Not reported
Harkanen et al, Nov 2019[39]	SAS Text miner	Hybrid	Not reported	72,390	Not reported
Hatef et al, 2019[45]	Not reported	Rule-based	LOINC, SNOMED	9,066,508	Not reported
Huang et al, 2021[44]	Not reported	Rule-based	Not reported	Not directly reported but averages reported (average of 9.6 notes per patient) than physicians (average of 6.5 notes per patient).	Not applicable
Hyun et al, 2009[43]	MedLEE	Rule-based	Not reported	553	Not reported
Hyun et al, 2020[15]	MedLEE	Rule-based	Not reported	553	Not reported
Karhade et al, 2021[40]	XGBoost	Rule-based	Not reported	Not reported	The area under the receiver operating curve (AUROC) of NLP algorithms for prediction of 90-day readmission using discharge summary notes, operative notes, nursing notes, physical therapy notes, case management notes, MD/APP notes were 0.70, 0.57, 0.57,

					0.60, 0.60, and 0.49, respectively.
Koleck et al, May/June 2021[42]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	UMLS, SNOMED CT	5,577,794	Recall ranged from .81 to .99, precision ranged from .75 to .96, and F1 ranged from .80 to .96, all indicating good or excellent system performance.
Koleck et al, Sept. 2021[14]	NimbleMiner	Hybrid (Neural Word Embeddings)	UMLS, SNOMED CT	504,395	Not reported
Korach et al, 2019[16]	TextRank and NC-Value	Hybrid (principal component analysis)	SNOMED-CT	778,956	The method achieved an average precision of 0.590 to 0.764 (when excluding numeric tokens). Time-dependent covariates Cox model using the phrases achieved a concordance index of 0.739. Clustering the phrases revealed clinical concepts significantly associated with RR event hazard.
Lehman et al, 2012[41]	Not reported	Hybrid (Hierchical Dirichlet Processes)	SNOMED, UMLS, RXNORM	Not reported	Not applicable
Marafino et al, 2015[38]	Not reported	Statistical (Stochastic Gradient Descent)	None	101,806	Not applicable
Nakayama et al, 2019[37]	Python (Gensim, Natural Language Toolkit); WordNet (TextBlob); Pattern for Python and Valence Aware Dictionary and sEntiment Reasoner (VADER)	Rule-based	ICD-9	not reported	Not reported

Neamatullah et al, 2008[36]	Not reported	Rule-based	SNOMED, UMLS	4,270	Performance evaluation of the de-identification software on the development corpus yielded an overall recall of 0.967, precision value of 0.749, and fallout value of approximately 0.002. On the test corpus, a total of 90 instances of false negatives were found, or 27 per 100,000-word count, with an estimated recall of 0.943. Only one full date and one age over 89 were missed.
Popejoy et al, 2015[30]	Not reported	Rule-based	Omaha system, UMLS, ICNP	128,135	Not reported
Press et al, 2015[35]	Not reported	Rule-based	Not reported	12,847	Identification of failed communication attempts using NLP had high external validity (kappa = 0.850, P < .001).
Song et al, 2021[13]	NimbleMiner	Hybrid (Neural Word Embeddings)	UMLS, SNOMED CT	2,610,757	NLP system's accuracy in identifying wound-related concepts in the clinical notes was high (>80%).
Sterling et al, 2019[34]	Not reported	Hybrid (Latent Dirichlet Allocation, Principal component analysis)	Not reported	256,878	Not applicable
Sterling et al, 2020[33]	Keras with TensorFlow 2.0 via Amazon Web Services Sagemaker	Rule-based	Not reported	226,317	Overall model accuracy and macro F1 score for number of resources were 66.5% and 0.601, respectively. The model had similar macro F1 (0.589 vs 0.592),

					and overall accuracy (65.9% vs 69.0%) compared to human raters.
Topaz et al, 2016[32]	MTERMS	Rule-based	SNOMED-CT, ICD-9, LOINC	1460	The overall system performance was good (F-measure is a compiled measure of system's accuracy = 92.7%), with best results for wound treatment (F-measure = 95.7%) and poorest results for wound size (F-measure = 81.9%).
Topaz et al, 2017[31]	MTERMS	Rule-based	SNOMED-CT, ICD-9, ICNP	202	The system achieved high accuracy: Precision = 95%; recall = 79.2%; F-measure = 86.4%; accuracy = 98%.
Topaz et al, Jan 2019[27]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	SNOMED-CT, ICD-9 and ICD10, ICNP	1,149,586	The overall F- score was 85.8% compared to 81% by the rule based-system with the best performance for identifying general fall history (F=89% vs. F=85.1% rule-based), followed by fall risk (F=87% vs. F=78.7% rule-based), fall prevention interventions (F=88.1% vs. F=78.2% rule-based) and fall within 2 days of the note date (F=83.1% vs. F=80.6% rule-based). The rule-based system achieved slightly better performance for fall within 2 weeks of the note date (F=81.9% vs. F=84% rule-based).

Topaz et al, Aug 2019[29]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	SNOMED-CT, ICD-10, ICNP, UMLS	463,544	NimbleMiner slightly outperformed other state-of-the-art NLP systems (average F-score = .84), while requiring significantly less time for the algorithms development.
Topaz et al, Nov 2019[12]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	SNOMED-CT, ICD-10, ICNP, UMLS	1,149,586	Models with larger word window width sizes (n = 10) that present users with about 50 top potentially similar terms for each (true) term validated by the user were most effective.
Topaz et al, Nov-Dec 2020[26]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	Not reported	727,676	Best performing text mining method (random forest) achieved good predictive performance of F-score=.82.
Topaz et al, 2021[25]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	SNOMED-CT, ICD-10, ICNP, UMLS	2,610,757	Overall, the NLP system achieved high symptom identification accuracy (F = 0.87) when tested on the gold standard human reviewed set of 500 clinical notes
Travers et al, 2003[60]	Not reported	Rule-based	UMLS	13,494	We found that 86% (4369) of the 5083 matched entries were identified with one UMLS concept only, and 14% were identified with two or more UMLS concepts. Accuracy was 92% confirmed through manual review.
Travers et al, 2013[58]	EMT-C (Emergency Medical Text Classifier)	Rule-based	Not reported	3353	In the manual classification of the 500 records by the two subject matter experts,

					we found 86.4 % agreement with a kappa of 0.55 (95% CI 0.47-0.64). EMT-C had a sensitivity, specificity, PPV (Positive Predictive Value), and NPV (Negative Predictive Value) of 0.79,0.83, 0.53 and 0.94 respectively which were all higher than baseline assessments.
Waudby-Smith et al, 2018[24]	TextBlob	Rule-based	Not reported	not reported	Not applicable
Woo et al, May 2021[22]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	UMLS (SNOMED-CT, ICD, ICNP	2610757	The NLP system achieved incredibly good overall performance (F measure = 0.9, 95% CI: 0.87–0.93) based on the test results obtained by using the notes for patients admitted to the ED or hospital due to UTI (Urinary Tract Infection).
Woo et al, June 2021[19]	NimbleMiner	Hybrid (Neural Word Embeddings, Random Forest)	UMLS	2610757	The natural language processing algorithm achieved overall good performance (F-measure = 0.88).
Yin et al, 2021[20]	latent Dirichlet allocation (LDA) in the Gensim Python package (version 3.8.0); Clinical Language Annotation, Modeling, and Processing software system	Hybrid (Latent Dirichlet Allocation)	Not reported	209,055	Not reported
Zhang et al, 2019[21]	R 3.3.2. with Principal component analysis	Hybrid (Logistic Regression and Principal	Not reported	not reported	Not reported

		Component analysis)			
Zhou et al, 2019[18]	StanfordCoreNLP	Statistical	Not reported	301	F1 of 81.1% for filtering out irrelevant information. Performance differences between this system and its baseline were statistically significant (P<.001; Wilcoxon test).