Supplemental Digital Content

Supplemental Methods

Ethical statement and IRB approvals:

The Beth Israel Deaconess Medical Center (BIDMC) and Massachusetts Institute of Technology (MIT) Institutional Review Boards (IRB) approved use of their de-identified dataset with external researchers following data-use agreements. The use of UCSF data was approved by the University of California, San Francisco (UCSF) institutional review board (IRB 19-29429). Procedures were followed in accordance with ethical standards of the responsible committee on human experimentation and with the Helsinki Declaration of 1975. No direct patient interactions were performed and all data was kept in secured digital locations behind a UCSF firewall (for UCSF data). We adhere to the TRIPOD reporting guidelines (see end of Supplement).

Overall statistical framework:

This study is to develop a theory of provider sentiment in clinical notes, to create a measurement approach for that theory and to perform initial validation of the measurement approach. In order to be reasonably comprehensive, we created two competing measurement methods for clinical sentiment: 1) one using a keyword dictionary and a simple proportion metric, which is an unsupervised approach and therefore not vulnerable to overfitting, and 2) a supervised approach using one of the highest performing deep learning models, which we hypothesized could achieve greater accuracy for measuring clinical sentiment compared to keywords at the cost of reduced interpretability and the need to control for potential overfitting.

Theorization and conceptualization:

We began our conceptualization of sentiment (used interchangeably with "opinion") from a well-accepted definition by Bing Liu: "An opinion is a quintuple, (e_i, a_{ij}, s_{ijkl}, h_k, t_l), where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k".[1] We further simplified this non-symbolically as: *judgement(s) made by a person (or author) directed at an object or target about an aspect of a more general entity*. For this proof-of-concept project, we further limited the previous statement to judgements made by an author regarding a general prognosis or direction of improvement or worsening of a particular problem, result or intervention. While sophisticated approaches to aspect-based sentiment analysis (or distinguishing targets of sentiment within a larger statement, sentence or document and their impact on overall sentiment) have been described elsewhere,[2] our initial approach was to demonstrate the possibilities of a domain-specific lexicon using opinionated keywords without incorporating the target (considered aspects and entities) of those opinions. In the future, following this proof-of-concept approach, more granular and aspect-oriented sentiment (whereby the target of sentiment is incorporated into our algorithms) will be pursued. An example sentence with labels from our dataset is shown in **eFigure 2** which is adapted from Weissman et al.[3]

General approach of global sentiment as proof-of-concept:

By using explicit keywords and demonstrating predictive validity, we propose a proof-of-concept that sentiment in the medical domain is possible and potentially useful. Our goals were to: (1) determine whether global subjective content could be approximated by using keywords representing negativity and positivity; (2) utilize an advanced deep learning tool to determine whether more sophisticated algorithms could improve sentiment accuracy over a keyword-based approach. Future work would continue with validation through the incorporation of additional labelers, qualitative surveys and interviews to perform additional validation and potentially incorporate aspects or targets of opinion to determine if additional feature extraction methods could enhance model prediction. All analyses using our two domain-specific sentiment methods were replicated on a more recent (2018-2019) but smaller dataset (n=1,123 patients) from UCSF.

Data Sources and Study Populations:

We used data from the MIMIC-III and UCSF deidentified datasets for this study. MIMIC-III is publicly available and all code and annotations used and made for this project can be made available on a public repository (GitHub). Replication code is available at https://github.com/ck37/mimic-clinical-sentiment. Sentiment analysis software is available at https://github.com/ck37/clinsent for general usage. However, deidentified UCSF data cannot be made publicly available per UCSF policy. Hence, we used MIMIC-III data for our primary analysis given it is more easily reproducible for readers/reviewers.

Details on UCSF De-identified data:

The UCSF de-identified dataset has been described in detail elsewhere.[9] Protected health information was removed in an automated fashion using the Philter algorithm that utilizes rule-based and statistical NLP. The authors acknowledge the use of the UCSF Information Commons computational research platform, developed and supported by UCSF Bakar Computational Health Sciences Institute. De-identified research data assets were used and available through UCSF Data Access for Research (University of California, San Francisco, Academic Research systems [2022]. UCSF DeID CDW-OMOP. 2022-June. University of California, San Francisco. Dataset. Available through https://ucsf.service-now.com/ucsfit?id=ucsf_sc_cat_item&sys_id=5d5fdd2cdbec3c908a57034b8a9619c8, which is a restricted permission environment).

Keyword generation:

Three clinical experts in critical care (JMC, AC, CC with backgrounds in Surgery, Critical Care, Palliative Care, Internal Medicine, Anesthesiology) participated in generation of positive and negative keywords until consensus was achieved. We prompted clinical experts by asking: ‘Which words or phrases represent positivity or negativity when conveying your thinking in clinical note in regard to a clinical problem, sign/symptom, therapy or prognosis?’. Examples of hypothetical statements were provided to help guide the keyword generation: ‘The patient is recovering well’ and ‘I am concerned about the patient’s prognoses.’ Discussion and debate of keywords were performed through individual meetings and shared digital documents (i.e. Google drive) allowing for multiple participants to share ideas and thoughts, individually then collectively. This was then performed iteratively until consensus was achieved on a medical lexicon of 72 positive and 103 negative shown in **eTable 1**.

This lexicon was broadened to include variations of keywords including alternate parts of speech, adverb and infinitive forms, contractions, and different tenses. Keywords were searched in each sentence or note fragment. Negation modifiers (i.e. “not” in front of a keyword) were distinguished from the positive form (or non-negated keyword) in this code by treating the positive form as a nested result, and were thus discarded. For example, 3 keywords from our lexicon naively match the excerpt “respiratory distress was not improving” – "distress”, “not improving”, and “improving”. Because “improving” was nested within a longer keyword phrase (“not improving”) it was not included as a match for this specific excerpt.

Manual Sentiment Labeling:

The purpose of the manual labeling was to create a “gold standard” reference allowing us to (1) validate a keyword-based lexicon for sentiment (i.e. does a note fragment containing a sentiment keyword *actually* describe subjective opinion according to a reader?); and (2) to train and test a state-of-the art algorithm that can identify and classify sentiment without using pre-defined keywords. While keywords are commonly used across most notes (**Table 1**), individual notes typically contain fewer than 10 keywords and depending on the note type, usually only 1-5 keywords (**eFigure 3**). If notes fragments are randomly sampled, then most fragments will not contain sentiment. Hence, we used a purposive approach to identify note fragments for labeling and training. To train our DeBERTa classifier, we used notes that contained at least one keyword, a form of study enrichment designed to improve statistical power. We extracted 1,493 note fragments that each contained at least one of the predefined keywords. These were then manually labeled using the prompt: "What overall sentiment or opinion is present in this note excerpt?". Sentiment was then labeled using a Likert-style rating scale of *very positive*, *positive*, *neutral*, *negative*, or *very negative*. For further validation, we performed manual labeling across 3 independent labelers in a subset of 100 note fragments (see "Preliminary Sentiment Measure Validation" under Materials and Methods in the main manuscript). Agreement improved across labelers when these categories collapsed into a 3-point scale (see manuscript results). In general, it is a good practice to record responses at a higher granularity than may be needed because it is always possible to aggregate them to a broader level of granularity, whereas if we had started with a 3-valued response we would not have been able to compare to the 5-valued results. Given ambiguity in the "very" categories and improved agreement with a 3-point scale, this was adopted for validation of the sentiment measure. For validation of both sentiment algorithms, we used all MIMIC notes and stratified notes into 5 equal parts: four parts by quartiles of sentiment based on percentage of negative keywords (keyword-based sentiment score) and a fifth part containing note segments that did not have a keyword.

Supervised Learning Approach for Sentiment Score Measure:

A preliminary labeled dataset of 1,493 MIMIC note excerpts was used to train a deep learning-based sentiment classification model. Sentences using the raw excerpt text without modification were used to train the model. The dataset was partitioned into a training set with 80% of the excerpts (1,194 observations) and a test set with the remaining 20% (299). The test set was excluded from model training. We used the Hugging Face Transformers framework[5] to fine-tune a DeBERTa-v3 pretrained model.[6] DeBERTa-v3 was chosen as the deep learning architecture due to its competitive performance in the leading natural language processing algorithm rankings: GLUE and SuperGLUE.[7,8]The model was fine-tuned for 5 epochs on the training set using categorical cross-entropy loss, a batch size of 16 observations, linear warmup of 50 steps, and AdamW optimizer with default settings (learning rate = 0.001, weight decay = 0). Opposite the keyword sentiment score, a higher DeBERTa-v3 sentiment score meant a higher positive sentiment. The keyword-based model was unsupervised, i.e. did not involve any labeled note excerpts, and therefore the sample splitting of the note excerpts was not relevant to it.

Data Missingness and Repeated/Redundant Notes:

For clinical outcome prediction, this was a complete case analysis and no imputation methods were performed. Some note types (e.g. nursing, rehabilitation) were found to be iterated from the previous day's note and thus produced text repeats in MIMIC. These were recognized during qualitative checks and we did not use other existing algorithms for 'note bloat' identification.[9] Only unique note elements for a particular day were used for analysis and model development.

Supplemental Tables

eTable 1-Sentiment Keywords Chosen for the Sentiment Lexicon

Negative Keywords	Positive Keywords
Abuse	Able
Bad	Advocate
Catastrophic	Agreeable
Challenging	Amenable
Concern	Appropriate
Concerned	Appropriately
Concerning	As expected
Declining	Better
Decompensated	Comfortable
Decompensating	Compensated
Despite	Compensating
Difficult	Controlled
Disappointing	Does want
Discouraged	Encouraged
Discouraging	Encouraging
Disseminated	Enjoy
Distress	Enjoyable
Distressed	Excellent
Distressing	Favorable
Does not enjoy	Fortunate
Does not want	Good
Doesn't want	Grateful
Eventful	Great
Failed	Improved
Frail	Improvement
Frustrating	Improving
Futile	In no distress
Getting worse	Looking forward
Grave	Low risk
Grim	Lower risk
Guarded	No concern
High risk	Not concerned
Higher risk	Not concerning
Hostile	Not discouraged
In distress	Not discouraging
Increased risk	Not eventful
Inoperable	Not getting worse
Instability	Not worse
Labile	Not worsening
Maximum	Operable
No improvement	Optimistic
No resolution	Optimized
Non operable	Outpatient
Not a candidate	Peaceful
Not able	Pleasant
Not amenable	Preferable
Not appropriate	Prefers
Not compensated	Properly
Not controlled	Realistic
Not encouraged	Reasonable
Not encouraging	Resolution
Not enjoyable	Resolved
Not favorable	Routine

Not fortunate	Satisfied
Not improved	Satisfies
Not improving	Satisfying
Not operable	Stability
Not preferable	Stable
Not realistic	Straightforward
Not reasonable	Successful
Not resolved	Superb
Not satisfied	Treated
Not satisfying	Unconcerned
Not stable	Unconcerning
Not treated	Uneventful
Not well controlled	Well compensated
Not well treated	Well controlled
Not within goals	Well treated
Not worthwhile	Within goals
Pessimistic	Wonderful
Poor	Worthwhile
Poorly compensated	Would want
Poorly controlled	
Poorly treated	
Refractory	
Risky	
Severe	
Unable	
Unappropriate	
Unappropriately	
Unclear	
Uncompensated	
Uncontrollable	
Uncontrolled	
Unfavorable	
Unfortunate	
Unimproved	
Unimproving	
Unknown	
Unoperable	
Unpreferable	
Unrealistic	
Unreasonable	
Unresolved	
Unsatisfied	
Unsatisfying	
Unstable	
Untreated	
Worrisome	
Worse	
Worsening	
Would not want	
Wouldn't want	

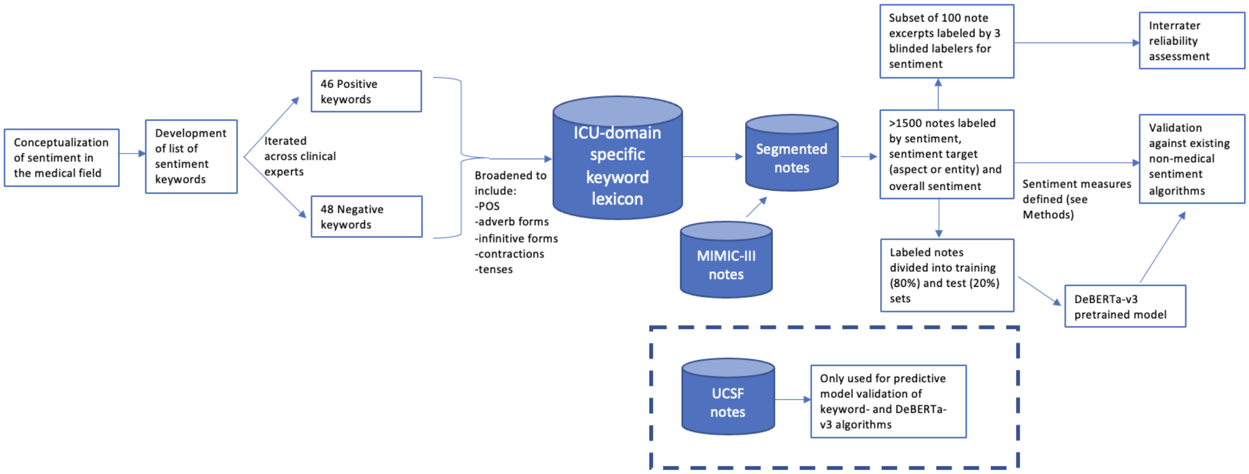
A medical lexicon of 72 positive and 103 negative keywords across clinical experts from diverse and multiple specialty backgrounds (General Surgery, Critical Care Medicine, Palliative Medicine, Internal Medicine, Anesthesiology). This list was broadened to include variations including alternative parts of speech, adverb forms, infinitive forms, contractions, and different tenses. Negations of the above were also accounted for in our analyses.

eTable 2: Detection of Sentiment Keywords by Note Category-

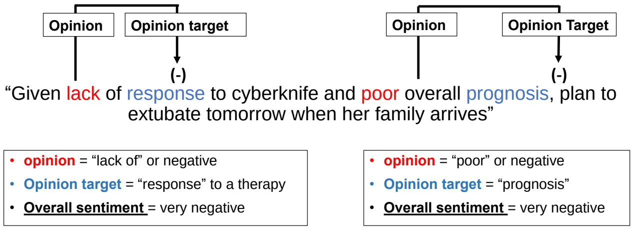
Note category	Notes (% of all notes of sample)	Percent with positive keyword	Percent with negative keyword	Percent with any sentiment keyword
Nursing	57,663 (28.8%)	67.6%	53.0%	81.1%
Physician	52,033 (26.0%)	89.9%	92.3%	97.8%
Respiratory	18,479 (9.2%)	18.7%	80.5%	85.0%
General	5,527 (2.8%)	50.3%	46.6%	66.3%
Nutrition	4,625 (2.3%)	32.8%	50.3%	65.9%
Rehab Services	4,581 (2.3%)	78.5%	63.1%	89.5%
Consult	69 (0.0%)	68.1%	82.6%	91.3%

Prevalence of use of any positive or negative keyword was calculated across pre-specified MIMIC-III note categories. Only 69 notes were categorized as Consult notes in MIMIC-III and likely labeled within Physician or General note categories.

Supplemental Figures


Conceptualization of sentiment in the medical field
Development of list of sentiment keywords
Iterated across clinical experts
46 Positive keywords
48 Negative keywords
Broadened to include:
-POS
-adverb forms
-infinitive forms
-contractions
-tenses
ICU-domain specific keyword lexicon
MIMIC-III notes
Segmented notes
Subset of 100 note excerpts labeled by 3 blinded labelers for sentiment
Interrater reliability assessment
>1500 notes labeled by sentiment, sentiment target (aspect or entity) and overall sentiment
Sentiment measures defined (see Methods)
Validation against existing non-medical sentiment algorithms
Labeled notes divided into training (80%) and test (20%) sets
DeBERTa-v3 pretrained model
UCSF notes
Only used for predictive model validation of keyword- and DeBERTa-v3 algorithms

eFigure 1: Cohort Selection and Study Design
This study included: (a) the development and validation of a sentiment measure. First, an ICU domain-specific lexicon (keyword dictionary) was constructed following conceptualization of sentiment and development of a list of positive and negative keywords by a broad multi-specialty group. Notes were then extracted from MIMIC-III and segmented. Of these, 1,493 note fragments were manually labeled for overall sentiment. Sentiment measures (including the primary measure of negative keyword frequency) were defined and a deep learning was also used to ensure at least one measure was built from a sophisticated methodology. These measures were then used for validation of various sentiment measures and existing non-medical sentiment algorithms (including Stanza, Sentimentr, Pattern).
Abbreviations: MIMIC-III = Multiparameter Intelligent Monitoring of Intensive Care III; UCSF = University of California, San Francisco; ICU=intensive care unit



“Given lack of response to cyberknife and poor overall prognosis, plan to extubate tomorrow when her family arrives”

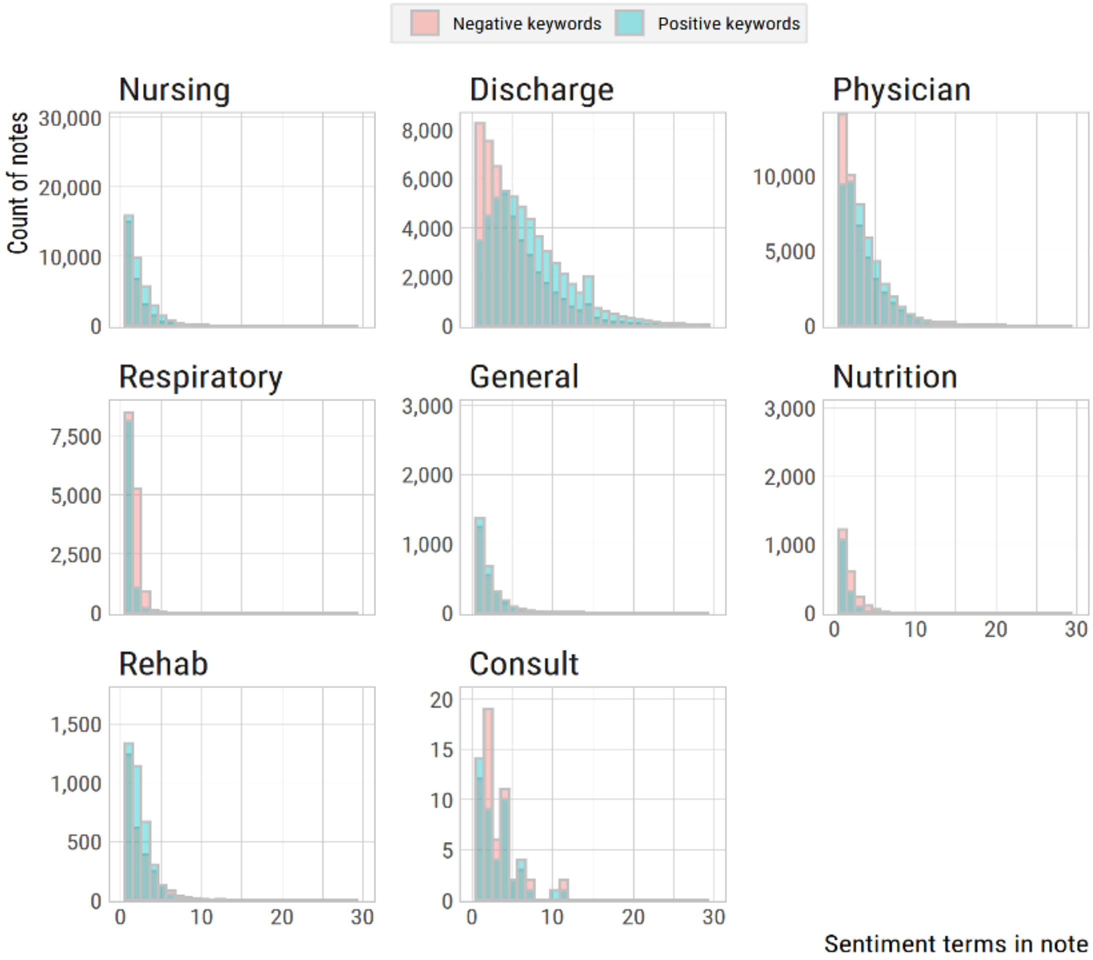
- **opinion** = “lack of” or negative
- **Opinion target** = “response” to a therapy
- **Overall sentiment** = very negative

- **opinion** = “poor” or negative
- **Opinion target** = “prognosis”
- **Overall sentiment** = very negative

eFigure 2: Example of Sentiment in an ICU Note Fragment

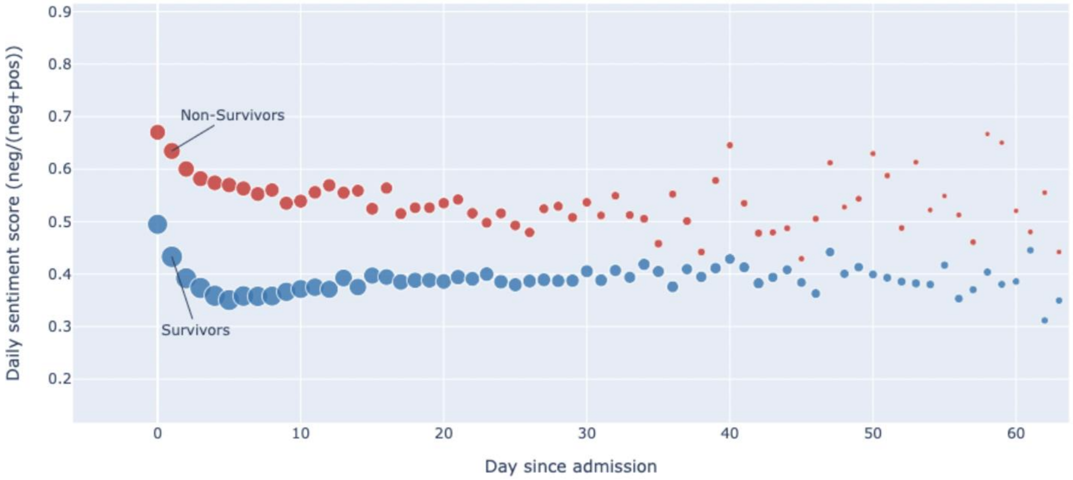
This example represents a sentence fragment in MIMIC-III. Following theorization of sentiment in the clinical domain, sentiment was labeled using a 5-point Likert scale including very positive, positive, neutral, negative and very negative. This 5-point scale was collapsed to 3-point (positive, neutral and negative) for the final regression analyses. Setniment targets were labeled as well as the global or overall sentiment of the fragment. In the left box, negative sentiment was inferred due to an outcome (“response” to a therapeutic) that was intrinsically deemed unsuccessful (by the ”lack of response”). The right box demonstrates a trajectory (or “prognosis”) that is worsening or unfavorable (”poor”) and negative sentiment was inferred. This figure was adapted from Weissman et al[3] and Liu [10]

Abbreviations: ICU=intensive care unit; MIMIC-III = Multiparameter Intelligent Monitoring of Intensive Care III



eFigure 3: Sentiment Terms Per Note by Category These distributions are from MIMIC. The mean number of keywords per note stratified by positive and negative keywords are presented. Each subfigure plots the mean number of keywords by note category on the x-axis across total number of notes on the y-axis. Notably, few notes are labeled as “consult” notes. It is likely that most notes written by consultations are embedded within “physician” and “general” note categories. Blue represents positive keywords and red, negative keywords.

Abbreviations: MIMIC-III = Multiparameter Intelligent Monitoring of Intensive Care III



eFigure 4- Keyword-based sentiment for ICU survivors and non-survivors as a function of time in the MIMIC dataset

Patients who survived (red) versus did not survive (blue) their ICU stay were stratified by ICU LOS (or time to death if in the ICU). Keyword-based sentiment scores were calculated for each stratification of patient by LOS and were the median of all notes per LOS stratification. A higher sentiment score represents increased negativity. Size of red or blue dots represents patient population with a given LOS. Fewer patients have a LOS beyond 20-30 days and hence, dots become smaller as a function of LOS. Abbreviations: ICU = intensive care unit; LOS = length of stay; MIMIC-III = Multiparameter Intelligent Monitoring of Intensive Care III;

Annotated TRIPOD Checklist:

Section/Topic	ı	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted. -Please refer to the Title on page 1 – this is a ‘development’ and proof-of-concept study	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. -Please refer to the “Measurements and Main Results” section of the abstract	4-5
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models. -The medical context can be found in the first 2 paragraphs of page 6 of the introduction	6
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both. -Please refer to the first paragraph of page 7	7
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable. -Retrospective cohort study; exploratory study	7, 9
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up. -2001-2012 dataset and note/clinical data spans the entire ICU stay -Validation on 2018-2019 UCSF dataset spanning the entire ICU stay	7
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres. -5 ICUs through MIMIC dataset (which is sourced from Beth Israel Deaconness Medical Center) -5 ICUs within UCSF hospital system	7-8
	5b	Describe eligibility criteria for participants. -Discharge summaries were excluded for ICU outcome analyses as these notes occurred after the outcome of interest but were included for sentiment prediction and descriptive results -We described inclusions of >=18 year old patients with at least one ICU note from a prespecified list of note types	8
	5c	Give details of treatments received, if relevant.	n/a
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed. -n/a	10-11, Supp
	6b	Report any actions to blind assessment of the outcome to be predicted. -not relevant because predictors sourced from EHR data (i.e. a retrospective study using existing MIMIC data) and this may be more relevant for a validation of an existing model in contrast to the development of the current one. This could be relevant for the DeBERTa model whereby the outcome is the chart review label of sentiment. Chart reviewers were blinded to the DeBERTa model itself. -For the subset of notes that were manually annotated across different annotators, annotations were done in a blinded fashion	n/a
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured. -For sentiment predictors, the sentence with raw excerpt text was used to train the model	10-11
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors. -For the subset of notes that were manually annotated across different annotators, annotations were done in a blinded fashion	n/a
Sample size	8	Explain how the study size was arrived at. -For the MIMC and UCSF datasets, all available observations were used. For the DeBERTa model, we collected pilot data based on availability of chart reviewers.	n/a

		Using transfer learning, we took advantage of a larger initial corpus to train the DeBERTa model. This pilot study will help produce power calculations for a future study (https://pubmed.ncbi.nlm.nih.gov/34461211/)	
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method. -Please see the data missingness section in the Supplement	Supp
Statistical analysis methods	10c	For validation, describe how the predictions were calculated. -Predictive validity was not performed for this study	11
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models. -For performance assessment, we used Spearman correlations	10-11
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done. -While validation of predictions were performed on an external dataset, this was not explicitly a validation study and thus no recalibration of our initial NLP sentiment models were performed	n/a
Risk groups	11	Provide details on how risk groups were created, if done. -n/a	n/a
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors. -This was primarily a development study but we did validate our predictions on an external smaller sample (1 years worth) of UCSF notes. We used the same eligibility criteria, outcomes and predictors for the external validation. The key differences between UCSF notes and MIMIC-III dataset is that we used a UCSF sample from 2018-2019 (MIMIC data is from 2001-2012) and patients have different baseline sociodemographic and likely clinical makeups given the geographic and historical differences of the respective institutions (UCSF vs. Beth Israel Deaconess).	n/a
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. -Refer to the CONSORT diagram in eFigure 1a-b for the flow of participants and data -Flow of participants and numbers with and without outcomes (with appropriate patient characteristics summarized) are found on page 12-13 (and Table 2)	12-13, supp
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome. -Characteristics of the cohort are found on page 12-13	12-13
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome). -Table 1	Table 1
Model performance	16	Report performance measures (with CIs) for the prediction model. -We report concordance and convergence across our sentiment classifiers compared to labelled note fragments and each other on page 12-13 -Interrater reliability with CIs for a subsample of notes is presented on page 13 -Predictive validity testing is on page 13-15 with CIs	12-14
Model-updating	17	If done, report the results from any model updating (i.e., model specification, model performance). -not relevant given that this was not a primary validation study; external UCSF dataset was used to replicate our models on more recent and an external dataset. No model updating was performed from the validation	n/a
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data). -Full limitations can be found from page 16-17. This includes labeling, ontologic, corpus and dataset limitations. It also highlights the need for further validation as this is an exploratory study.	16
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data. -page 15, eTable 3	15, etable 3

	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence. -see Page 14-18 in the discussion for complete interpretation, limitations and relevance to other studies	14-18
Implications	20	Discuss the potential clinical use of the model and implications for future research. -Refer to beginning and end of discussion for future clinical use (this is proof-of-concept and not for clinical use in its current form) and for future research implications	14,16
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets. -Replication code is available at https://github.com/ck37/mimic-clinical-sentiment. Sentiment analysis software is available at https://github.com/ck37/clinsent for general usage.	19
Funding	22	Give the source of funding and the role of the funders for the present study. -Funding sources: Dr. Cobert was supported by the UCSF Initiative for Digital Transformation in Computational Biology & Health, the Hellman Fellows Foundation and supported by the UCSF Claude D. Pepper Older Americans Independence Center funded by NIA (P30 AG044281). Dr. Lee was supported by the National Institute of Aging K24AG066998 and R01AG057751. Dr. Smith was supported by grants from the NIA (R01AG057751 and K24AG068312). Dr. Pirracchio is supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award Center of Excellence in Regulatory Science and innovation grant to University of California, San Francisco (UCSF) and Stanford University, U01FD005978. -Funding role: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication	2

REFERENCES:

1. Liu B. *Sentiment Analysis and Opinion Mining*. Morgan&Claypool Publishers; 2012:1-184.
2. Lu HY, Yang J, Hu C, Fang W. One for "All": a unified model for fine-grained sentiment analysis under three tasks. *PeerJ Comput Sci*. 2021;7:e816. doi:10.7717/peerj-cs.816
3. Weissman GE, Ungar LH, Harhay MO, Courtright KR, Halpern SD. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *J Biomed Inform*. 01 2019;89:114-121. doi:10.1016/j.jbi.2018.12.001
4. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc*. 01 01 2018;25(1):32-39. doi:10.1093/jamia/ocx084
5. Wolf T, Debut L, Sanh V, et al. Transformers: State-of-the-Art Natural Language Processing. Association for Computational Linguistics; Oct 2020:
6. He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. 2021;arXiv:2111.09543
7. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv*. 2018;doi:arXiv:1804.07461
8. Wang A, Pruksachatkun Y, Nangia N, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*. 2019;doi:arXiv:1905.00537
9. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and Sources of Duplicate Information in the Electronic Medical Record. *JAMA Netw Open*. 09 01 2022;5(9):e2233348. doi:10.1001/jamanetworkopen.2022.33348
10. Liu B. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers; 2012:184.