HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2020 January 01.

Published in final edited form as:

J Biomed Inform. 2019 January ; 89: 114–121. doi:10.1016/j.jbi.2018.12.001.

Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness

Gary E. Weissman[a,b,c,*], **Lyle H. Ungar**[d], **Michael O. Harhay**[b,e], **Katherine R. Courtright**[a,b,c], and **Scott D. Halpern**[a,b,c,e]

[a]Pulmonary, Allergy, and Critical Care Division, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[b]Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[c]Leonard Davis Institute of Health Economics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[d]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

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

Abstract

Sentiment analysis may offer insights into patient outcomes through the subjective expressions made by clinicians in the text of encounter notes. We analyzed the predictive, concurrent, convergent, and content validity of six sentiment methods in a sample of 793,725 multidisciplinary clinical notes among 41,283 hospitalizations associated with an intensive care unit stay. None of these approaches improved early prediction of in-hospital mortality using logistic regression models, but did improve both discrimination and calibration when using random forests. Additionally, positive sentiment measured by the CoreNLP (OR 0.04, 95% CI 0.002–0.55), Pattern (OR 0.09, 95% CI 0.04–0.17), sentimentr (OR 0.37, 95% CI 0.25–0.63), and Opinion (OR 0.25, 95% CI 0.07–0.89) methods were inversely associated with death on the concurrent day after adjustment for demographic characteristics and illness severity. Median daily lexical coverage ranged from 5.4% to 20.1%. While sentiment between all methods was positively correlated, their agreement was weak. Sentiment analysis holds promise for clinical applications but will require a novel domain-specific method applicable to clinical text.

[*]Corresponding author at: Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, 306 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA. gary.weissman@uphs.upenn.edu (G.E. Weissman).

Conflict of interest

The authors declared that there is no conflict of interest.

Keywords

Critical care; Forecasting; Natural language processing; Electronic health records; Attitude of health personnel

1. Introduction

In the era of widespread adoption of electronic health records (EHRs) [1] and learning health systems [2] there is growing interest in improving the utilization of free-text data sources. Among patients with critical illness, the text of clinical notes has been used to identify diagnoses and interventions in the intensive care unit (ICU) and to improve predictions of future health states [3–6]. Clinical text contains important diagnostic information not found in structured data sources within the EHR [7,8]. But clinicians also make subjective assessments [9] and express attitudes about patient outcomes that may be purposefully or unwittingly inscribed in clinical notes. It is unknown if analysis of these subjective attitudes may augment existing yet imperfect mortality predictions [10], improve communication by highlighting affective dynamics underlying patient-provider and patient-surrogate relationships [11], or provide a feedback mechanism to clinicians regarding their implicit biases [12].

The study of attitudes expressed in text is called “sentiment analysis” or “opinion mining” [13]. Dictionaries of terms (i.e. lexica) containing words with associated sentiment vary across different domains [14]. For example, “soft” may imply a different sentiment whether used with respect to sports or toys [15]. The analysis of sentiment in a medical context has been limited to patient opinions expressed in online social media [16,17] and in suicide notes [18], the association of sentiment in hospital discharge documents [19] and nursing notes [20] with mortality, and a descriptive comparison between nursing and radiology notes [21].

Therefore, we sought to determine the construct validity of existing sentiment methods derived from other domains when used for analysis of clinical text among patients with critical illness. Specifically, we examined the predictive, concurrent, content, and convergent validity of these methods to assess different aspects of the sentiment construct.

2. Materials and methods

2.1. Population and data source

We analyzed the Medical Information Mart for Intensive Care (MIMIC) III database which comprises all hospital admissions requiring ICU care at the Beth Israel Deaconess Medical Center in Boston, MA, between 2001 and 2012 [22]. Only hospital admissions with at least one clinical encounter note and a length of stay (LOS) ≤ 30 days were included.

2.2. Text sources and sentiment methods

We aggregated clinical encounter notes at the patient-day level for each hospital admission and included notes from physicians, nurses, respiratory therapists, and other clinical

specialties. We calculated the proportion of positive sentiment in each collection of daily aggregated notes as

$$Proportion\ of\ positive\ sentiment = \frac{\sum_w P_w}{\sum_w P_w + \sum_w |N_w|},$$

where P_w and N_w are the positive and negative sentiment scores, respectively, for each word w in the daily aggregated text. The same approach was used for sentence-level sentiment results. We calculated separate scores using the Opinion [23], AFINN [24], EmoLex [25], Pattern [26], sentimentr [27], and the Stanford deep learning sentiment model [28] implemented in the CoreNLP [29] toolkit. The first five methods use simple dictionary lookups. The sentimentr and Pattern methods use dictionary lookups and also account for valence shifters (e.g. “very” and “not”). The CoreNLP method uses a deep learning model trained with phrase-level sentiment labels in parsed sentences and thus identifies sentiment at the sentence level.

2.3. Construct validity

The validity of a construct or instrument is determined by how well it measures some true phenomenon of interest [30]. We sought to determine how well sentiment – as defined in numerous sentiment analysis methods [23–29] that have been developed in non-clinical domains – captures actual clinician sentiment expressed in the text of clinical encounter notes. For each sentiment method, we examined different aspects of validity individually in order to make an overall assessment of its validity in the clinical domain.

2.3.1. Predictive validity—A sentiment measure with predictive validity should be strongly associated with some future outcome [31]. Therefore, for each sentiment method, we trained a logistic regression model based on a random 75% sample of all hospital admissions to predict in-hospital mortality using data from the first day of the hospitalization. Logistic regression was chosen for its ease in reproducibility and interpretation [32]. The proportion of positive sentiment on the first hospital calendar day was included as a feature, and each model was adjusted for age, gender, initial ICU type, modified Elixhauser score [33,34], and initial sequential organ failure assessment (SOFA) score [35]. The same input variables were used to build a set of random forest models [36] to account for potential non-linear decision boundaries and complex interactions between input variables that would not be captured in the logistic regression model [37]. For the random forest models, the number of variables to consider at each split in a tree was determined by maximizing the classification accuracy with 10-fold cross-validation [38,39]. Each model was compared to a baseline model with the same clinical and demographic covariates but that did not incorporate any sentiment measure. Model discrimination was assessed with the C-statistic which is equivalent to the area under the curve of the receiver operating characteristic and measures the discrimination of a binary classifier [40]. Comparisons of C-statistics were made with the DeLong method which is a non-parametric test based on the theory of U-statistics and which accounts for the correlated nature of predictions from the same data [41]. Calibration was assessed with the Brier score which is a strictly proper scoring rule that describes the error of continuous risk predictions of a binary classifier

[42,43] Comparisons were made using a bootstrapped [44] *t*-test with 1000 replicates. All performance measures were reported using the remaining 25% hold-out testing sample which would allow for sufficient power to detect a meaningful difference of 0.03 in the C-statistic [45].

2.3.2. Concurrent validity—A measure with concurrent validity should be strongly associated with an outcome that is measured in the same time period [31]. This is in contrast to predictive validity which requires the association of a current observation with a future outcome. Therefore, we examined the relationship between sentiment and the risk of mortality on the same day. We constructed a multivariable mixed-effects logistic regression model using the daily proportion of positive sentiment as the primary, time-varying exposure and daily risk of in-hospital death as the dichotomous outcome. The model was adjusted for age, gender, initial ICU type, and modified Elixhauser score [33,34]. A random effect was included for each hospital admission to account for repeated observations. A SOFA score ≥ 7 was included as a dichotomous, time-varying exposure to account for daily changes in clinical severity. While daily SOFA scores have not been studied with respect to the daily risk of death, a time-varying score of ≥ 7 has been associated with an approximately 20% mortality rate in the ICU [46].

2.3.3. Convergent validity—A measure with convergent validity should agree with other measures that describe the same phenomenon. This is critical for assessing the validity of sentiment measures because the object toward which sentiment is directed (e.g. the patient, the prognosis, the tumor) may vary significantly. Each lexicon may also vary by the content of its terms and associated sentiment depending on the domain in which the method was developed and original purpose [21]. Therefore, each sentiment method may provide a measure of some different aspect of the complex tapestry of sentiment found in clinical encounter notes. To assess the degree to which these six sentiment methods described the same phenomena, i.e. their convergence [47], we measured their agreement with Cronbach's alpha and calculated pair-wise Pearson correlations (r) at the patient-day level. Given there was no clear comparison group from which these sentiment methods should diverge, we did not also assess discriminant validity.

2.3.4. Content validity—A useful construct of sentiment in clinical encounter notes should rely on keywords commonly used in the medical domain. Thus, the content validity is the extent to which a sentiment approach is capable of accounting for words and phrases found in these texts [31]. We measured this lexical coverage as the proportion of words in each patient-day's aggregated text sample that was found in the lexicon. Because the CoreNLP method implements a pre-trained deep learning model, we used all unique tokens from the original training set to identify a maximum upper bound on lexical coverage.

2.4. Non-mortal outcomes

In order to identify other potential relationships with sentiment measures, we assessed the correlation between sentiment and two nonmortal outcomes. First, we measured the correlation between the daily proportion of positive sentiment and the mean self-reported pain score among subjects for whom numerical pain scores were recorded in the nursing

flowsheet. Second, we measured the correlation between the proportion of positive sentiment averaged over the entire hospital stay and the total hospital length of stay in days. Both correlations were measured using Pearson's method as described above.

Mixed-effects regression models were built using Stata (version 14.2, StataCorp, College Station, TX). Extraction of sentiment and training of other models were performed with the R language for statistical computing (version 3.3.2). The Pattern sentiment method was implemented using the Python programming language (version 2.7.13). The Stanford CoreNLP toolkit (version 3.9.1) was run using Java (version 8). We used a two-sided alpha = 0.05 as a threshold for significance and adjusted all tests for multiple comparisons (Bonferroni correction). This study was determined to be exempt by the Institutional Review Board of the University of Pennsylvania.

3. Results

We analyzed 41,283 unique hospital admissions comprising 331,972 patient-days. The median hospital LOS was 6 days (Interquartile range [IQR] 4–8), the median age at admission was 61 years (IQR 40–76), and 4033 (9.8%) patients died in the hospital. Each hospital admission contained a median of 8 (IQR 4–21) clinical encounter notes with median 1438 words (IQR 573 – 5,169). These totaled 793,725 encounter notes containing 229,037,446 words (Fig. 1). The distribution of daily sentiment for each method is summarized in Table 1.

The unadjusted temporal trajectories of sentiment stratified by inhospital mortality are presented in Fig. 2. However, the baseline model and all logistic regression models with the addition of sentiment had C-statistic 0.81 without clinically relevant differences in discrimination ($p = 0.026–0.948$ for all comparisons). There were no meaningful differences in calibration with the addition of sentiment to a baseline model (all models had Brier score 0.074; $p = 0.083–0.847$). In contrast, the random forest models all increased the baseline C-statistic from 0.95 to 0.97 or 0.98 ($p < 0.001$ for all comparisons). Similarly, the addition of sentiment to random forest models improved the Brier score from 0.074 to 0.067 or 0.065 ($p < 0.001$ for all comparisons) across all sentiment measures.

Sentiment was strongly associated with death when measured on the concurrent day for four of the six sentiment methods (Table 1). After adjustment for baseline characteristics and daily severity of illness, the proportion of positive sentiment measured by the CoreNLP method was inversely associated with the daily risk of death (OR 0.04, 95% CI 0.002–0.55).

As a measure of convergence, the Cronbach's alpha for sentiment estimates for each patient-day was 0.65 (95% CI 0.64–0.65). All correlations between methods were positive and statistically significant, but most were of a modest magnitude ($p < 0.001$; Fig. 3). The median proportion of daily lexical coverage by hospital admission (Fig. 4) ranged from 5.4% to 20.1% among those methods using a lexicon-based approach.

The most common terms from the Opinion lexicon and representative samples of text are presented in Table 2. The associated polarity of these terms included instances with both concordant and discordant meanings in the medical domain.

Numeric values for patient-reported pain were available for 19,199 hospitalizations across 48,104 patient-days. Among patient-days with recorded values, there were median 5 (IQR 2–9) pain observations per patient-day. Daily sentiment was not strongly correlated with the mean daily self-reported pain (Fig. 5). The daily proportion of positive sentiment averaged over the entire hospitalization was weakly correlated with hospital length of stay.

4. Discussion

In our assessment of multidisciplinary encounter notes of patients hospitalized with critical illness, existing sentiment approaches demonstrated little evidence of most types of validity and exhibited high variability between methods. These results argue against the use of available sentiment methods to inform bedside clinical decisions, but also highlight opportunities to make sentiment methods more clinically applicable.

Many of the covered terms in this analysis had discordant polarity when applied in the medical domain. For example, the term “right” in medical parlance most often expresses anatomic laterality (e.g. “right ventricle”), and thus should carry a neutral rather than positive sentiment with respect to prognosis or clinical condition. Similarly, the term “bs” is a shorthand abbreviation with multiple senses and may indicate “breath sounds”, “bowel sounds”, or “blood sugar” depending on the context. It should carry a neutral valence for all of these medical uses, but carried a negative polarity in the Opinion lexicon, where it may have been used originally to indicate a vulgar term in the online consumer reviews of electronics products.

The strong concurrent validity after adjustment for clinical and demographic characteristics suggests a temporal sensitivity of sentiment to the patient’s clinical condition on the same day. This finding was true even with adjustment for changes in severity of illness on each day, highlighting the presence of additional information encoded in free-text data not found in structured data sources such as laboratory values and vital signs. The models with the strongest effect sizes (i.e. lowest odds ratios) in this analysis (CoreNLP, Pattern, and sentimentr; Table 1) were the only three that accounted for varying degrees of context in contrast to the other methods that used simple dictionary lookups. Nuances in expression of clinician sentiment are likely better captured by these approaches.

However, the addition of sentiment measures to a baseline prediction model resulted in no meaningful improvements to its discrimination or calibration using logistic regression models. In contrast, the addition of sentiment to random forest models uniformly improved both discrimination and calibration by a small amount. This finding suggests that interactions between sentiment measures and other features may yield predictive information with respect to mortality predictions. While we hypothesize that some severity information not captured in standard risk scores (e.g. SOFA) may be encoded in the sentiment of clinical notes, it is unclear how predictive performance would change with domain-appropriate word sense and improved lexical coverage.

Although all sentiment estimates were positively correlated with each other, their overall agreement was poor. The Opinion, AFINN, and EmoLex approaches were more highly

correlated with each other (r = 0.58–0.68), while the Pattern and sentimentr approaches were weakly correlated (r = 0.33). These latter two methods were only very weakly correlated with the CoreNLP method (r = 0.09), despite the strong concurrent validity observed among these three. These findings suggest a weak convergence towards two or three distinct constructs. More work is needed to distinguish between the sources, objects, and aspects of sentiment in clinical text. Additionally, the sentiment associated with objective medical terms (e.g. “cardiac arrest”) is distinct from the expression of a private state [48] of a clinician (e.g. “Mr. Jones is an unpleasant and uncooperative 65 year old man”). Each of these has separate analytic and ethical implications for use in clinical predictive modeling that have yet to be explored.

Daily and aggregated sentiment were not consistently correlated with patient-reported pain or hospital length of stay, respectively. More positive sentiment was, in most cases, paradoxically, weakly correlated with more pain. It is unclear if patient pain itself may induce positive affect in clinicians via sympathy, or if sympathetic clinicians are at baseline more inclined to record higher pain scores and write more positive notes. The variable concordance between length of stay and pain associations suggests a large degree of noise in measurements using these methods from non-medical domains.

Finally, the content of sentiment lexica demonstrated coverage of medical terms that was higher than in previous analyses of medical text, but low compared to sentiment use in other domains. For example, Denecke et al. found coverage of 5–11% in radiology reports, 6–11% in discharge summaries, and 8–12% in nursing notes, depending on the sentiment lexicon [21]. Coverage for the widely used SemEval Dataset range from 8% to 89% percent using commonly available sentiment lexica [49]. The use of deep learning models like that in the CoreNLP toolkit highlights a challenge in their assessment compared to other models that use a transparent and human-readable lexicon.

The results of this study should be interpreted in the context of some limitations. First, the study analyzed data from a single academic center and may not generalize to the documentation style or patient population in other settings. Second, our analysis did not distinguish between the emotional valence of objective and subjective terms which conflates their practical use in clinical risk prediction. Third, the divergent results in predictive validity across model types does not definitively address the optimal model specification for using sentiment for mortality predictions. Fourth, we only analyzed the daily proportion of positive sentiment and did not explore other aggregate measures such as maximum, minimum, or measures of central tendency, that might be informative. Fifth, the results presented here maybe sensitive to different pre-processing methods for sentence boundary detection, word tokenization, and other steps that were not explored [50]. Finally, differences in sentiment between clinical specialties, discipline, and location may exhibit important variation and were not examined in this study.

4.1. Conclusions

In conclusion, this is the first study to examine sentiment in a set of multidisciplinary clinical encounter notes of critically ill patients and to assess the validity of these measures. Sentiment is strongly and concurrently associated with the risk of death even after

adjustment for baseline characteristics and severity of illness. Our findings highlight the need for a domain-specific sentiment lexicon that has wide coverage of medical terminology with appropriate word senses, and that accounts for negation, intensifiers, and temporal relations. Any medical sentiment method, because it may be used for high-stakes clinical decision making, should also balance the needs for performance (e.g. complex annotation pipelines as inputs to deep learning models) and interpretability (e.g. lexicon, n-gram, and other bag-of-word methods). Future work should seek to validate these findings in a broader population, better distinguish sources and objects of sentiment, and address potential ethical challenges of using sentiment to guide clinical care.

Acknowledgments

Funding

GEW received support from the National Institutes of Health (T32-HL098054, K23-HL141639). MOH received support from the National Institutes of Health (K99-HL141678).

References

[1]. Henry J, Pylypchuk Y, Searcy T, Patel V, Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015, The Office of National Coordinator for Health Information Technology, 2016.

[2]. Krumholz HM, Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system, Health Aff. (Milwood) 33 (2014) 1163–1170

[3]. Weissman GE, Harhay MO, Lugo RM, Fuchs BD, Halpern SD, Mikkelsen ME, Natural language processing to assess documentation of features of critical illness in discharge documents of acute respiratory distress syndrome survivors, Ann. Am. Thorac. Soc 13 (2016) 1538–1545. [PubMed: 27333269]

[4]. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ, Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis, J. Am. Med. Inform. Assoc (2014) 801–807. [PubMed: 24384230]

[5]. 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 54 (2015) 114–120. [PubMed: 25700665]

[6]. Lehman LW, Saeed M, Long W, Lee J, Mark R, Risk stratification of ICU patients using topic models inferred from unstructured progress notes, AMIA Annu. Symp. Proc 2012 (2012) 505–511. [PubMed: 23304322]

[7]. Zhou L, Baughman AW, Lei VJ, Lai KH, Navathe AS, Chang F, Sordo M,Topaz M, Zhong F, Murrali M, Navathe S, Rocha RA, Identifying patients with depression using free-text clinical documents, Stud. Health Technol. Inform 216 (2015) 629–633. [PubMed: 26262127]

[8]. Navathe AS, Zhong F, Lei VJ, Chang FY, Sordo M, Topaz M, Navathe SB, Rocha RA, Zhou L, Hospital readmission and social risk factors identified from physician notes, Health Serv. Res 53 (2018) 1110–1136. [PubMed: 28295260]

[9]. Detsky ME, Harhay MO, Bayard DF, Delman AM, Buehler AE, Kent SA,Ciuffetelli IV, Cooney E, Gabler NB, Ratcliffe SJ, et al., Discriminative accuracy of physician and nurse predictions for survival and functional outcomes 6 months after an ICU admission, JAMA 317 (2017) 2187–2195. [PubMed: 28528347]

[10]. Sinuff T, Adhikari NK, Cook DJ, Schünemann HJ, Griffith LE, Rocker G, Walter SD, Mortality predictions in the intensive care unit: comparing physicians with scoring systems, Crit. Care Med 34 (2006) 878–885. [PubMed: 16505667]

[11]. Jacobowski NL, Girard TD, Mulder JA, Ely EW, Communication in critical care: family rounds in the intensive care unit, Am. J. Crit. Care 19 (2010) 421–430. [PubMed: 20810417]

[12]. Chapman EN, Kaatz A, Carnes M, Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities, J. Gen. Int. Med 28 (2013) 1504–1510.

[13]. Liu B, Sentiment analysis and subjectivity, Handbook of Natural Language Processing, 2010, pp. 627–666.

[14]. Andreevskaia A, Bergler S, When specialists and generalists work together: Overcoming domain dependence in sentiment tagging, in: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 290–298.

[15]. Hamilton WL, Clark K, Leskovec J, Jurafsky D, Inducing domain-specific sentiment lexicons from unlabeled corpora, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, 2016 (2016) 595–605.

[16]. Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, Gonzalez GH, Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts, J. Biomed. Inform 62 (2016) 148–158. [PubMed: 27363901]

[17]. Ji X, Chun SA, Wei Z, Geller J, Twitter sentiment classification for measuring public health concerns, Social Network Anal. Min 5 (2015) 13.

[18]. Pestian JP, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J,Cohen KB, Hurdle J, Brew C, Sentiment analysis of suicide notes: a shared task, Biomed. Inform. Insights 5 (2012) 3–16. [PubMed: 22419877]

[19]. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH, Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study, PLOS ONE 10 (2015) 1–10.

[20]. Waudby-Smith IER, Tran N, Dubin JA, Lee J, Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients, PLOS ONE 13 (2018) e0198687. [PubMed: 29879201]

[21]. Denecke K, Deng Y, Sentiment analysis in medical settings: new opportunities and challenges, Artif. Intell. Med 64 (2015) 17–27. [PubMed: 25982909]

[22]. Johnson AEW, Pollard TJ, Shen L, Lehman L.-w.H., Feng M, Ghassemi M,Moody B, Szolovits P, Anthony Celi L, Mark RG, MIMIC-III, A freely accessible critical care database, Sci. Data 3 (2016).

[23]. Hu M, Liu B, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 168–177.

[24]. Årup Nielsen F, A new ANEW: evaluation of a word list for sentiment analysis in microblogs, in: Proceedings of the ESWC2011 Workshop on ‘Making Sense of Microposts’: Big things come in small packages, 2011.

[25]. Mohammad SM, Turney PD, Crowdsourcing a word-emotion association lexicon, Comput. Intell 29 (2013) 436–465.

[26]. Smedt TD, Daelemans W, Pattern for python J Mach. Learn. Res 13 (2012) 2063–2067.

[27]. Rinker TW, sentimentr: Calculate text polarity sentiment, University at Buffalo/SUNY, Buffalo, New York, 2017 http://github.com/trinker/sentimentr. Accessed December, 2017.

[28]. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C, Recursive deep models for semantic compositionality over a sentiment treebank, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1631–1642.

[29]. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D, The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60.

[30]. Bolarinwa OA, Principles and methods of validity and reliability testing of questionnaires used in social and health science researches, Niger. Postgrad. Med. J 22 (2015) 195–201. [PubMed: 26776330]

[31]. Cronbach L, Meehl P, Construct validity in psychological tests, Psychol. Bull 52 (1955) 281–302. [PubMed: 13245896]

[32]. Larsen K, Petersen JH, Budtz-Jørgensen E, Endahl L, Interpreting parameters in the logistic regression model with random effects, Biometrics 56 (2000) 909–914. [PubMed: 10985236]

[33]. Elixhauser A, Steiner C, Harris DR, Coffey RM, Comorbidity measures for use with administrative data, Med. Care 36 (1998) 8–27. [PubMed: 9431328]

[34]. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ, A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data, Med. Care 47 (2009) 626–633. [PubMed: 19433995]

[35]. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H,Reinhart CK, Suter PM, Thijs LG, The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine, Intensive Care Med. 22 (1996) 707–710. [PubMed: 8844239]

[36]. Liaw A, Wiener M, Classification and regression by randomForest, R News 2 (2002) 18–22.

[37]. Hastie T, Tibshirani R, Friedman J, The elements of statistical learning, Springer Series in Statistics, New York, NY, USA, 2008.

[38]. Kuhn M, Johnson K, Applied predictive modeling, Springer, 2013.

[39]. James G, Witten D, Hastie T, Tibshirani R, An introduction to statistical learning, Springer, 2013.

[40]. Tripepi G, Jager KJ, Dekker FW, Zoccali C, Statistical methods for the assessment of prognostic biomarkers (Part I): discrimination, Nephrol. Dial. Transplant 25 (2010) 1399–1401. [PubMed: 20139066]

[41]. DeLong ER, DeLong DM, Clarke-Pearson DL, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics (1988) 837–845. [PubMed: 3203132]

[42]. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW, Assessing the performance of prediction models: a framework for some traditional and novel measures, Epidemiology (Cambridge, Mass.) 21 (2010) 128.

[43]. Gneiting T, Raftery AE, Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc 102 (2007) 359–378.

[44]. Angelo Canty BD Ripley, boot: Bootstrap R (S-Plus) Functions, 2017.

[45]. Hajian-Tilaki K, Sample size estimation in diagnostic test studies of biomedical informatics, J. Biomed. Inform 48 (2014) 193–204. [PubMed: 24582925]

[46]. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL, Serial evaluation of the SOFA score to predict outcome in critically ill patients, JAMA 286 (2001) 1754–1758. [PubMed: 11594901]

[47]. Campbell DT, Fiske DW, Convergent and discriminant validation by the multi-trait-multimethod matrix, Psychol. Bull 56 (1959) 81–105. [PubMed: 13634291]

[48]. Wiebe J, Wilson T, Cardie C, Annotating expressions of opinions and emotions in language, Language Resour. Eval 39 (2005) 165–210.

[49]. Gatti L, Guerini M, Turchi M, SentiWords: deriving a high precision and high coverage lexicon for sentiment analysis, IEEE Trans. Affect. Comput 7 (2016) 409–421.

[50]. Park A, Hartzler AL, Huh J, McDonald DW, Pratt W, Automatically detecting failures in natural language processing tools for online community text, J. Med. Internet Res 17 (2015) e212. [PubMed: 26323337]

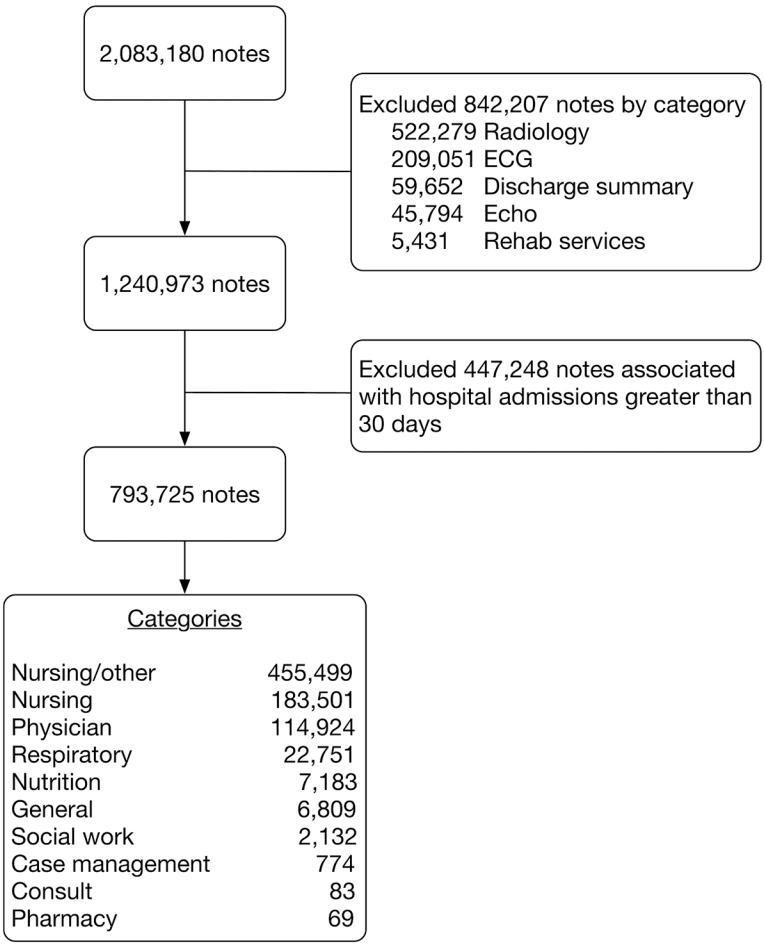

Fig. 1.
Exclusions of clinical notes used in the analytic sample. ECG = electrocardiogram.

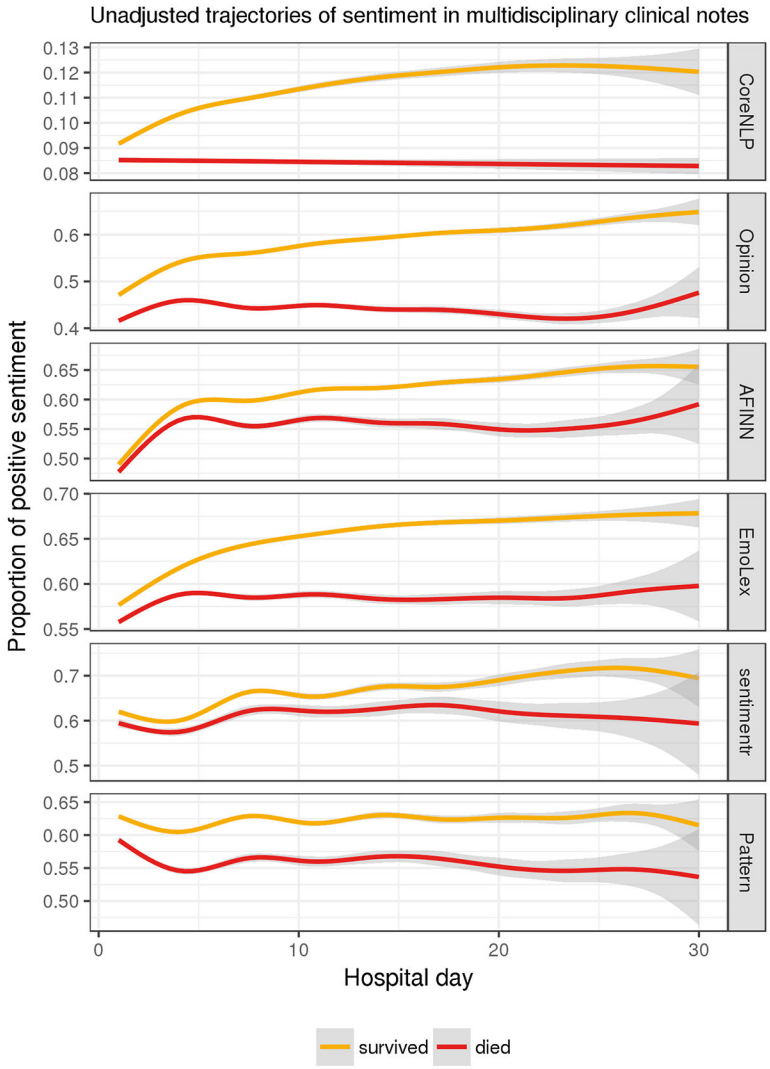

Fig. 2.

Unadjusted trajectories of the proportion of positive sentiment by sentiment method using a generalized additive model smoother with 95% confidence intervals. All sentiment trends demonstrated clear separation by survival status.

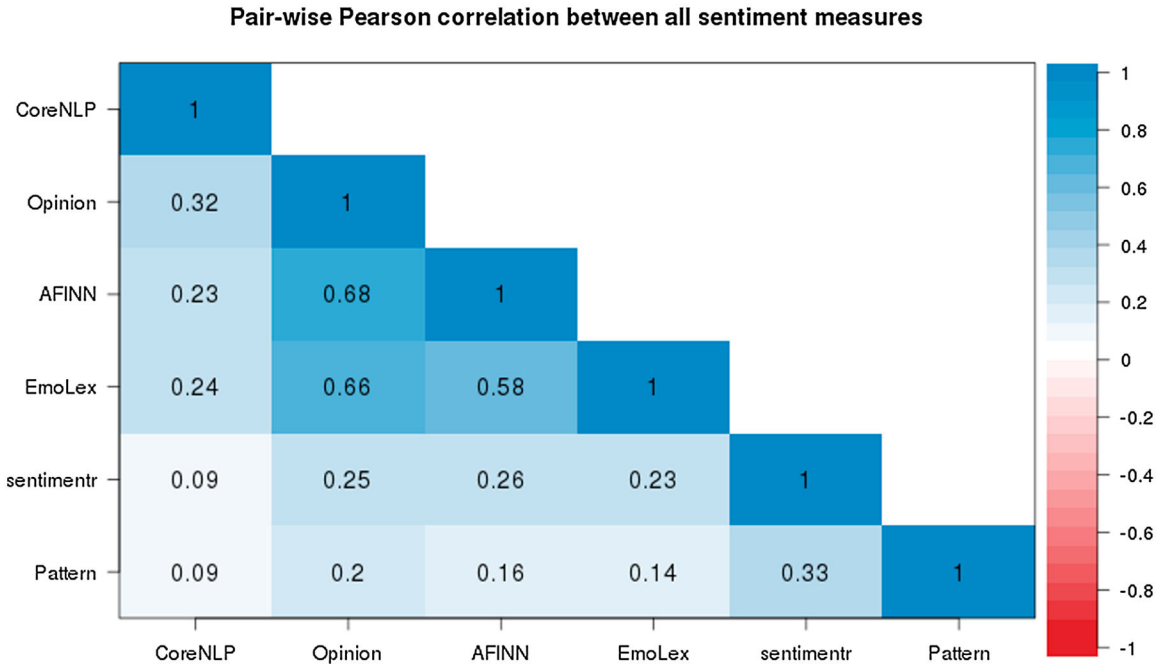

Fig. 3.

Pair-wise Pearson correlations between methods of calculated sentiment by patient-day. All estimates have $p < 0.001$ after adjustment for multiple comparisons. Correlations between sentiment methods are highly variable.

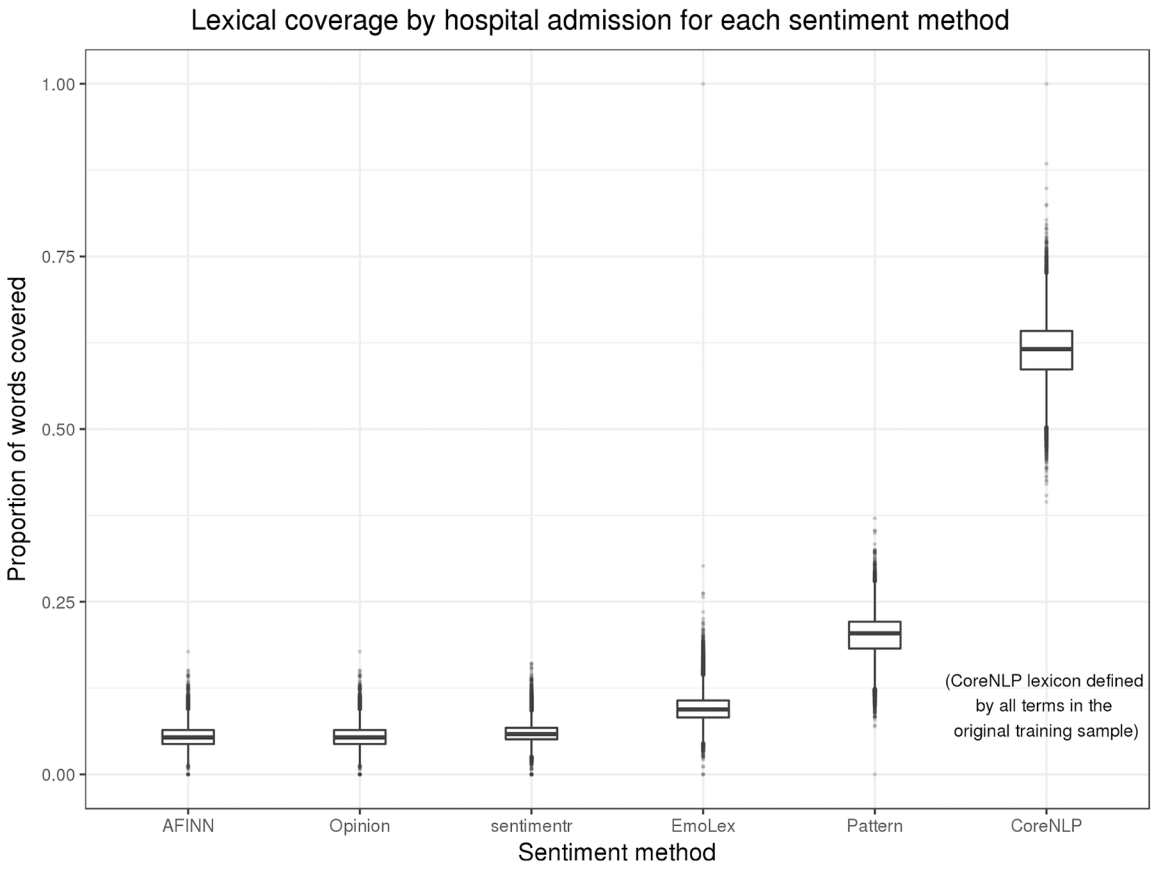

Fig. 4.

Distribution of the median proportion of covered words for each hospital admission by sentiment method. Most sentiment methods exhibited low lexical coverage in clinical notes. The CoreNLP method is not lexicon based, and so the estimate of coverage based on the training corpus may be overly optimistic.

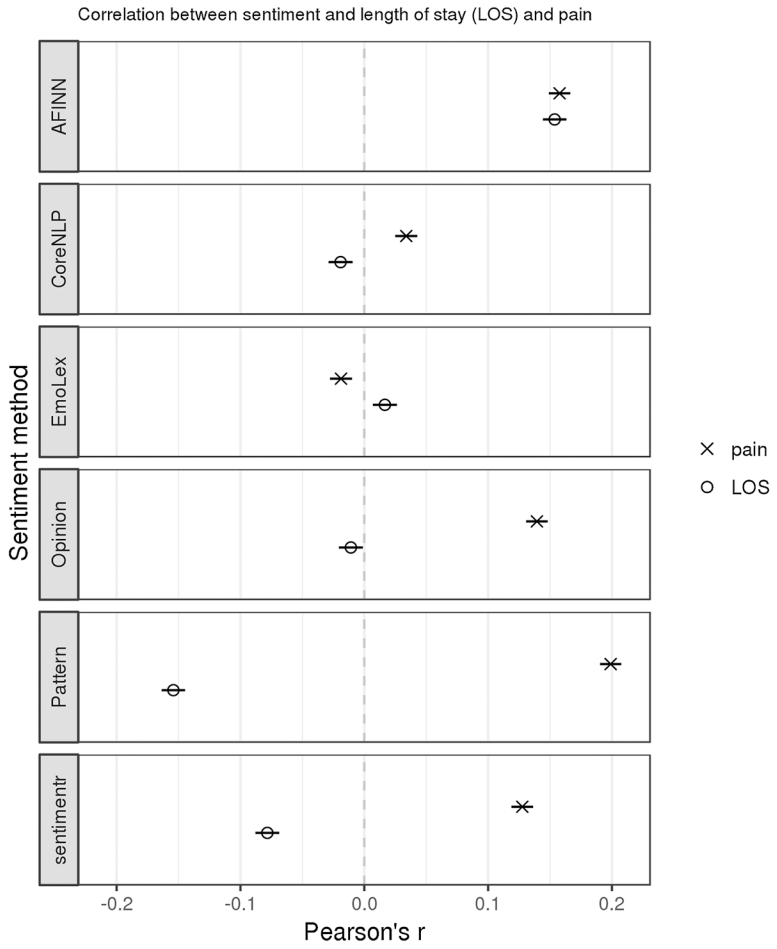

Fig. 5.

Correlation between sentiment methods and hospital length of stay and self-reported pain. Point estimates are reported using Pearson's r with bars indicating 95% confidence intervals. Sentiment measures exhibited wide variation with both concordant and discordant correlations of expected relationships.

Table 1

Adjusted odds ratio estimate for the proportion of daily positive sentiment associated with same-day mortality for each sentiment method based on mixedeffects logistic regression model to assess concurrent validity; and distribution of daily sentiment. CI = confidence interval. IQR = interquartile range.

Sentiment method	Odds ratio (95% CI)	p value	Median (IQR)
Opinion	0.25 (0.07–0.89)	0.033	0.50 (0.39–0.62)
EmoLex	1.89 (0.41–8.69)	0.412	0.60 (0.53–0.69)
AFINN	0.65 (0.23–1.87)	0.428	0.56 (0.44–0.68)
Pattern	0.09 (0.04–0.17)	< 0.001	0.63 (0.53–0.74)
sentimentr	0.37 (0.25–0.63)	< 0.001	0.71 (0.35–0.96)
CoreNLP	0.04 (0.002–0.55)	0.017	0.09 (0.05–0.14)

Table 2

The most common terms from the Opinion lexicon found in the clinical text sample, their defined sentiment, and the number of appearances of each term across all clinical notes. Representative examples from the text demonstrated both concordance and discordance with the calculated sentiment.

Term	Sentiment	Appearances (n)	Representative context
Pain	negative	658,808	‘Pt reports back pain’, ‘Continue to monitor pain’
Patient	positive	588,213	‘Encouraged patient to take his medicine’, ‘I saw and examined the patient’
Stable	positive	411,028	‘stable frontal infarct’, ‘remains hemodynamically stable’
Right	positive	383,482	‘only moving right arm’, ‘elevation of the right hemidiaphragm’
Clear	positive	368,261	‘w/o clear evidence of infiltrates’, ‘Nutrition: clear liquids, advance diet’
Well	positive	365,899	‘get radiation as well as this decision’, ‘satting well, no resp distress’
Support	positive	325,814	‘s/p arrest requiring ventilatory support’, ‘Emotional support given to patient & family’
Soft	positive	290,426	‘abdomen soft slightly distended’, ‘possibility of soft tissue pus collection’
Failure	negative	268,838	‘PNA with hypercarbic respiratory failure’, ‘R-sided heart failure leading to hepatopedal flow’
bs	negative	259,638	‘PULM: decreased bs on left’, ‘soft distended with hypoactive bs’