Identification of Serious Illness Conversations in Unstructured Clinical Notes using Deep Neural Networks

Isabel Chien[1], Alvin Shi[1], Alex Chan[2], and Charlotta Lindvall[3,4]

[1] Massachusetts Institute of Technology, Cambridge, USA
chieni@mit.edu, alvinshi@mit.edu
[2] Harvard T.H. Chan School of Public Health, Boston, USA
alexchan@mail.harvard.edu
[3] Dana-Farber Cancer Institute, Boston, USA
[4] Brigham and Women's Hospital, Boston, USA
clindvall@mail.harvard.edu

Abstract. Advance care planning, which includes clarifying and documenting goals of care and preferences for future care, is essential for achieving end-of-life care that is consistent with the preferences of dying patients and their families. Physicians document their communication about these preferences as unstructured free text in clinical notes; as a result, routine assessment of this quality indicator is time consuming and costly. In this study, we trained and validated a deep neural network to detect documentation of advanced care planning conversations in clinical notes from electronic health records. We assessed its performance against rigorous manual chart review and rule-based regular expressions. For detecting documentation of patient care preferences at the note level, the algorithm had high performance; F1-score of 92.0 (95% CI, 89.1-95.1), sensitivity of 93.5% (95% CI, 90.0%-98.0%), positive predictive value of 90.5% (95% CI, 86.4%-95.1%) and specificity of 91.0% (95% CI, 86.4%-95.3%) and consistently outperformed regular expression. Deep learning methods offer an efficient and scalable way to improve the visibility of documented serious illness conversations within electronic health record data, helping to better quality of care.

Keywords: deep learning, end-of-life care, palliative care, natural language processing, clinical notes, electronic health records

1 Introduction and Related Work

To ensure that patients receive care that is consistent with their goals, clinicians must communicate with seriously ill patients about their treatment preferences. More than 80% of Americans say they would prefer to die at home, if possible. Despite this, 60% of Americans die in acute care hospitals and 20% die in an Intensive Care Unit (ICU)[1]. Advance care planning, which includes clarifying and documenting goals of care and preferences for future care, is essential for achieving end-of-life care that is consistent with the preferences of seriously ill patients

and their families. Inadequate communication is associated with more aggressive care near the time of death, decreased use of hospice and increased anxiety and depression in surviving family members[2–5]. Several studies have demonstrated the potential of advanced care planning to improve end-of-life outcomes (e.g., reducing unintended ICU admissions and increasing hospice enrollment). In the absence of explicit goals of care decisions, clinicians may provide clinical care[6] that does not provide a meaningful benefit to the patient[7] and, in the worse case, interferes with the treatment of other patients[6]. For these reasons, it is recommended that care preferences are discussed and documented in the EHR within the first 48 hours of an ICU admission[8, 9].

In recent years a consensus has emerged that such conversations are an essential component of practice and must be monitored to improve care quality. However, the difficulty of retrieving documentation about these conversations from the electronic health record has limited rigorous research on the prevalence and quality of clinical communication. For example, the National Quality Forum (NQF) recommends that goals of care be discussed and documented in the EHR within the first 48 hours of an ICU admission, especially for frail and seriously ill patients. This was one of only two Centers for Medicare and Medicaid Services recommended palliative care quality measures for the Medicare Hospital Inpatient Quality Reporting program[10]. Yet, despite widespread support, routine assessment of this and similar quality measures have proven nearly impossible because the information is embedded as non-discrete free-text within clinical notes. Manual chart review is time-consuming and expensive to scale [11–13]. Consequently, many end-of-life quality metrics are simply not assessed, and their impact on distal and important patient outcomes have been insufficiently evaluated.

The emergence of omnipresent EHRs and powerful computers present novel opportunities to apply advanced computational methods such as natural language processing (NLP)[14] to assess end-of-life quality metrics including documentation of ACP. NLP enables machines to process or understand natural language in order to perform tasks like extracting communication quality embedded as non-discrete free-text within clinical notes[15].

Two main approaches to NLP information extraction exist. Rule-based extraction uses a pre-designed set of rules[14], which involves computing curated rules specified by experts, resulting in algorithms that detect specific words or phrases. This approach works well for smaller defined sets of data such as when searching for all the brand names of a generic medication (e.g., if X is present, then Y=1). However, rule-based approaches fail when the desired information appears in a large variety of contexts within the free text[16].

Recent advances in machine learning coupled with increasingly powerful computers have created an opportunity to apply advanced computational methods, such as deep learning, to assess the content of free-text documentation within clinical notes. Such approaches possess the potential to broaden the scope of research on serious illness communication, and when implemented in real-time, to change clinical practice.

In contrast to rule-based methods, deep learning does not depend upon predefined set of rules. Instead, these algorithms learn patterns from a labeled set of free-text notes and apply them to future datasets[16]. A deep learning-based approach works well for tasks for which the set of extraction rules is very large, unknown, or both. In deep learning, algorithms can learn feature representations that aid in interpreting varied language.

In this study, we used deep learning[17] to train models to detect documentation of serious illness conversations, and we assess the performance of these deep learning models against manual chart review and rule based regular expression.

2 Data

2.1 Data Source

We derived our sample from the publicly available ICU database, Multi Parameter Intelligent Monitoring of Intensive Care (MIMIC) III, developed by the Massachusetts Institute of Technology (MIT) Lab for Computational Physiology and Beth Israel Deaconess Medical Center (BIDMC)[18]. It is a repository of de-identified administrative, clinical, and survival outcome data from more than 58,000 ICU admissions at BIDMC from 2001 through 2012. Between 2008 and 2012, the dataset also included clinical notes associated with each ICU admission. The Institutional Review Board of the BIDMC and MIT have approved the use of the MIMIC-III database by any investigator who fulfills data-user requirements. The study was deemed exempt by the Partners Institutional Review Board.

2.2 Cohort

The study population included adult patients (age $\geq$18) who were admitted to the medical, surgical, coronary care, or cardiac surgery ICU. The training and validation set included physician notes from patients who died during the hospital admission to ensure that we would have sufficient examples of documentation of care preferences. We excluded patients who did not have physician notes within the first 48 hours because these patients either died shortly after admission or transferred out of the ICU.

2.3 Clinical domains

Our main outcome was to identify documentation of care preferences within 48 hours of an ICU admission in seriously ill patients. We aimed to detect the binary absence or presence of any clinical text that fit specified documentation of domains: patient care preferences (goals of care conversations or code status limitations), goals of care conversations, code status limitations, family communication (which included communication or attempt to communicate with family that did not result in documented care preferences), and full code status.

Domains were chosen by board-certified, experienced palliative care clinicians through a lengthy and iterative process. They determined categories that are both relevant to widespread existing palliative care quality measures and interesting to future research questions. The specifications of each domain are outlined (Table 1).

Table 1. Clinical domain specifications.

Domain	Documentation example
Patient care preferences	Fulfills criteria for goals of care conversations and/or code status limitations
Goals of care conversations	Explicitly shown preferences about the patients goals, values, or priorities for treatment and outcomes. Does NOT include presumed full code status or if obtained from other sources.
Code status limitations	Explicitly shown preference of patients care restricting the invasive care. Includes taken over preference from previous admission.
Communication with family	Explicit conversations held during ICU stay period with patients or family members about the patients goals, values, or priorities for treatment and outcomes.
Full code status	Explicitly or implicitly shown preference for full set of invasive care including intubation and resuscitation. Includes presumed full code status or if obtained from other sources.

2.4 Annotation

We developed a set of abstraction guidelines to ensure reliable abstraction between annotators. Each annotator identified clinical text that fit specified communication domains and labeled the portions of text identified for a domain, with no restrictions on length of a single annotation.

A gold standard dataset, considered to contain true positives and true negatives, was developed through manual annotation by a panel of four clinicians. Annotation was done using PyCCI, a clinical text annotation software developed by our team. Each note was annotated by at least two clinicians and annotations were then validated by a third clinician. Similar to previously published chart abstraction studies performed for this measure, the abstraction team had real-time access to a US board certified hospice and palliative medicine attending physician-expert reviewer, met weekly, and used a log to document common questions and answers to facilitate consistency[11, 19].

The clinician coders manually annotated an average of 239 notes each (SD, 196), for a total of 641 notes. Each note contained an average of 1397 tokens (IQR, 1004-1710). The inter-rater reliability among the four clinician annotators

was kappa > 0.65 at the note level for each domain. The performance of each clinician coder was varied–for example, they identified documentation of care preferences with a sensitivity ranging from 77-92% (in comparison to the final gold standard).

3 Methods

3.1 Pre-processing

Annotated notes were pre-processed for both rule-based regular expression and neural network methods. First, texts were cleaned to remove any extraneous spaces, lines, or characters. Each cleaned note was tokenized, which means it was split into identifiable elements–in this case, words and punctuation. We used the Python module spaCy in order to tokenize intelligently, based on the structure of the English language[20]. Labels were associated with individual tokens and datasets were split out by domain, as each method was run separately.

3.2 Regular expression

Our baseline model is a simple regular expression based on pre-curated rules for each domain. Appendix A shows the rules used for each domain. These rules are keywords that the regular expression program identifies as belonging to its corresponding domain, taking into account varieties in punctuation and case. To create the regular expression library, we identified tokens that were sensitive and specific for each prediction task. We calculated sensitivity by evaluating the proportion of a token's total number of occurrences that were labeled for each domain. We evaluated specificity by evaluating what proportion of a token's total number of occurrences were in a note that was in an unlabeled note for each domain. A board-certified clinician used these data points–sensitivity, specificity, frequency that each token appeared on the labeled text and frequency in texts outside of the domain–and their clinical knowledge to generate a list of terms that could likely be generalized.

Regular expressions identify patterns of characters exactly as they are specified in a set of rules. If text in the note matches a keyword in the regular expression library for the domain, it is labelled as positive for that concept. This method acts as a baseline to compare our algorithm against. We used a regular expression program, ClinicalRegex, also developed by our lab[30]. ClinicalRegex is easily accessible and intuitive to navigate, which makes it an efficient choice for groups that are not able to employ computer scientists. We have chosen to compare our deep learning methods against an easily understandable and accessible method to illustrate the benefits of more complex methods.

3.3 Artificial neural network

Deep learning involves training a neural network to learn data representation and fulfill a specified task. We trained algorithms to identify clinical text documentation of serious illness communication. During the training process, the

neural network learns to identify and categorize tokens (individual words and symbols) as belonging to each of the pre-specified domains and maximizes probability across predicted token labels[21].

The specific neural network used, NeuroNER, was developed by Dernoncourt et al. for the purpose of named-entity recognition[22]. NeuroNER has been evaluated for use in the de-identification of patient notes[21]. It allows for each token to be labelled only with a single label. However, tokens in our study were often associated with multiple labels. For example, a sentence could indicate that both communication with family occurred and that goals of care were discussed. In order to allow for multi-class labelling, a separate, independent model was trained per domain. For each domain, the data set was split up into randomized training and validation sets, with 70% (449 notes) of the set in training, and 30% (192 notes) in validation.

With the parameters derived from this training process, the model is run on the validation data set to examine its performance on a data set it was not specifically tuned to fit. Performance on the validation set also determines when training converges, indicating that the model is optimally trained. Training converges when there has been no improvement on the validation set performance in ten epochs. The neural network ultimately determines domain labels for each token. From the predicted token-level results, a note-level classification is determined by the presence or absence of labelled tokens by domain in each note. We used Tensorflow version 1.4.1 and trained our models on a NVIDIA Titan X Pascal GPU. Below are the hyperparameters selected for our use:

- character embedding dimension: 25
- character-based token embedding LSTM dimension: 25
- token embedding dimension: 100
- label prediction LSTM dimension: 100
- dropout probability: 0.5

For our experiments, we were able to compare our gold standard labels, derived from manual annotation by clinicians as described in Section 2.4, to the predicted output to evaluate the performance of the neural network and the regular expression method.

4 Results

4.1 Evaluation metrics

Algorithm performance was determined at two levels: token-level and note-level, referring to the binary absence or presence of a label at these levels. Token-level results are more specific and allow accurate identification of relevant text within clinical notes. Note-level results allow determination of whether documentation of communication occurred. At both of these levels, we calculated the following metrics: sensitivity, specificity, positive predictive value, accuracy, and F1-score. The F1-score is the harmonic average of positive predictive value and sensitivity.

It allows us to determine the success of our algorithm both in identifying true positives as well as true negatives.

The 95% confidence intervals for all metrics were determined via bootstrapping[23]; each trained network model was validated for 1,000 trials in addition to the reported performance point. During each trial, a validation set of 192 notes was created by random sampling with replacement of the original validation set of 192 unique notes. This creates an approximate distribution of performance for the model. In basic bootstrap technique, the 2.5th and 97.5th percentiles of the distributions for each metric are taken as the 95% confidence interval[24].

4.2 Performance

Table 2 summarizes the performance of the regular expression method and Table 3 summarizes the performance of the neural networks in identifying documentation of serious illness communication at the note level, for each clinical domain, on the validation set. Figure 1 displays a comparison in the F1-scores for each domain. For identification of documentation of patient care preferences, the algorithm achieved an F1-score of 92.0 (95% CI, 89.1-95.1), with 93.5% (95% CI, 90.0%-98.0%) sensitivity, 90.5% (95% CI, 86.4%-95.1%) positive predictive value and 91.0% (95% CI, 86.4%-95.3%) specificity. For identification of family communication without documentation of preferences, the algorithm achieved an F1-score of 0.91 (95% CI, 0.87-0.94), with 90.7% (95% CI, 86.0%-95.9%) sensitivity, 90.7% (95% CI, 86.5%-94.8%) positive predictive value and 92.5% (95% CI, 89.2%-97.8%) specificity. Token-level performance is displayed in Appendix B.

At the note-level, we have been able to achieve high accuracy for all domains and see that in the validation set, the neural network outperforms the regular expression method in every domain for F1-score, significantly so in identifying patient care preferences, goals of care conversations, and communication with family. These domains contain more complex and diverse language, which are successfully identified by the neural network. A static library is not able to capture the diversity in these domains, necessitating the use of machine learning.

4.3 Error analysis

A review of documentation that the neural networks identified as serious illness conversations that was not labeled serious illness conversations in the gold standard (false positives) showed that the algorithm identified documentation that clinician coders missed. Though our gold standard was rigorously reviewed and validated, there still remains room for human error. Comparing the identified text from the neural network and regular expression methods, we found that as expected, the neural network was able to identify complex and unique language that the regular expression method was not. Doctors employ diverse and non-standardized language in clinical notes; we require more flexible and extensible methods in order to efficiently process this information. Static libraries cannot capture the full complexity of language without sacrificing sensitivity or

Table 2. Performance (%) of the regular expression method on the validation data set.

Domain	F1-score	Accuracy	Sensitivity	Positive Predictive Value	Specificity
Patient care preferences	76.0	78.6	70.7	82.3	86.0
Goals of care conversations	37.2	57.8	26.1	64.9	87.0
Code status limitations	94.3	96.4	98.3	90.6	95.5
Communication with family	43.6	67.7	27.9	100.0	100.0
Full code status	90.9	88.5	84.6	98.2	96.8

Table 3. Performance (%) of the neural networks on the validation data set. Values in parentheses are 95% confidence intervals.

Domain	F1-score	Accuracy	Sensitivity	Positive Predictive Value	Specificity
Patient care preferences	92.0 (89.1-95.1)	92.2 (89.6-95.1)	93.5 (90.0-98.0)	90.5 (86.4-95.1)	91.0 (86.4-95.3)
Goals of care conversations	85.7 (80.4-90.3)	89.1 (85.6-92.4)	85.1 (78.4-91.5)	86.3 (80.0-93.0)	91.5 (87.7-95.7)
Code status limitations	95.9 (93.0-98.7)	97.4 (95.8-99.2)	98.3 (96.9-100.0)	93.5 (89.2-97.7)	97.0 (95.0-98.9)
Communication with family	90.7 (87.4-93.9)	91.7 (89.1-94.4)	90.7 (86.0-95.9)	90.7 (86.5-94.8)	92.5 (89.1-95.9)
Full code status	98.5 (97.5-99.4)	97.9 (96.6-99.2)	100.0 (100.0-100.0)	97.0 (95.1-98.9)	93.5 (89.2-97.7)

specificity–they must be curated such that library terms are not too broad and they are not able to utilize context. All note-level identification can be traced to the detection of specific words with examples of text for each method provided in Appendix C.

4.4 Effect of training set size

In order to determine how smaller training sets related to the performance of the trained algorithms, we trained multiple networks with varying number of notes. We plotted training dataset size against algorithm performance for 8 sample sizes (Figure 2). The performance seemed to plateau at around 200 notes (around 250,000 tokens), which suggests that annotation efforts can be efficiently leveraged to generalize the models to varied health systems.

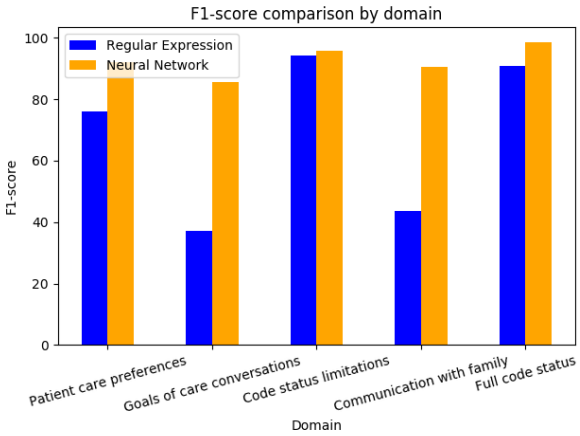

Fig. 1. Comparison between the F1-score of the regular expression method and neural networks by domain.

5 Discussion and future work

We describe a novel use of deep learning algorithms to rapidly and accurately identify documentation of serious illness conversations within clinical notes. When applied to identifying documentation of patient care preferences, our algorithm demonstrated high sensitivity (93.5%), positive predictive value (90.5%) and specificity (91.0%), with a F1-score of 92.0. In fact, we found that deep learning outperformed individual clinician coders both in terms of identifying the documentation and in terms of its many-thousands-time-faster speed.

Existing work has shown that machine learning can extract structured entities like medical problems, tests and treatments from clinical notes[25, 26], and unstructured image-based information in radiology, pathology and opthamology[27–29]. Our study extends this line of work and demonstrates that deep learning can also perform accurate automated text-based information classification.

Up until now, extracting goals of care documentation nested within free-text clinical notes has relied on labor-intensive and imperfect manual coding[11]. Using the capabilities of deep learning as demonstrated in this paper would allow for rapid audit and feedback regarding documentation at the system and individual practitioner level. This would result in significant opportunities for quality improvement that are currently not being met. Deep learning models could also improve patient care in real-time by broadening what is available at the point of care in the EHR. For example, clinicians could view displays of all documented goals of care conversations, or be prompted to complete documentation that was not yet available.

Important limitations must be noted. Deep learning algorithms only detect what is documented. It is not fully understood to what extent documentation reflects the actual content of a patient-clinician conversation surrounding serious

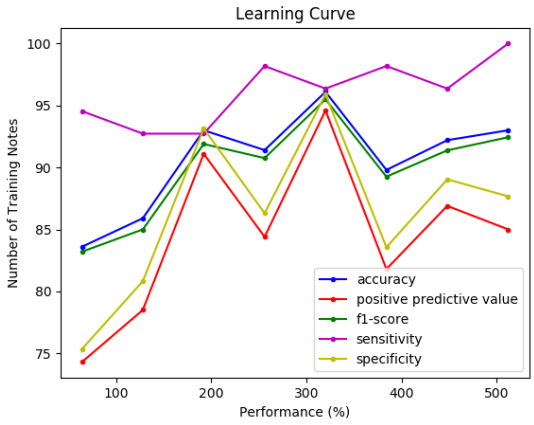

Fig. 2. Neural network performance on validation set for detection of note-level documentation of patient care preferences by number of notes used for training.

illness care goals. However, documentation is the best proxy we have to understand and to track these conversations. This is also a single institution study, which may limit its generalizability. Future work will involve the investigation of how extensible models are to clinical notes from different health system. Variations in EHR software and the structure of clinical notes in different institutions makes it essential to further train and validate our methods using data from multiple healthcare systems. This should be imminently possible, as our learning curve suggested that the neural network needed to train on as few as 200 clinician coded notes to perform well. Future research should also focus on optimizing deep neural networks to further improve performance, and on determining the feasibility of operationalizing this algorithm across institutions.

6 Conclusion

This is the first known report of employing deep learning, to our knowledge, to identify serious illness conversations. The potential of this technology to improve the visibility of documented goals of care conversations within the EHR and for quality improvement has far reaching implications. We hope such methods will become an important tool for evaluating and improving the quality of serious illness care from a population health perspective.

Acknowledgements

We are particularly grateful to Tristan Naumann, Franck Dernoncourt, Elena Sergeeva, Edward Moseley, and Alistair Johnson for helpful guidance and advice during the development of this research. Additionally, we would like to thank

Peter Szolovits for providing computing resources, as well as Saad Salman, Sarah Kaminar Bourland, Haruki Matsumoto and Dickson Lui for annotating clinical notes. This research was facilitated by preliminary work done as part of course HST.953 in the Harvard-MIT Division of Health Sciences and Technology (HST) at Massachusetts Institute of Technology (MIT), Boston, MA.

References

1. Cook, D., Rocker, G. Dying with Dignity in the Intensive Care Unit. N Engl J Med 2014; 370:2506-2514
2. Wright AA, Zhang B, Ray A, et al. Associations between end-of-life discussions, patient mental health, medical care near death, and caregiver bereavement adjustment. JAMA. 2008;300(14):1665-1673.
3. Nicholas LH, Langa KM, Iwashyna TJ, Weir DR. Regional variation in the association between advance directives and end-of-life Medicare expenditures. JAMA. 2011;306(13):1447-1453.
4. Teno JM, Gruneir A, Schwartz Z, Nanda A, Wetle T. Association between advance directives and quality of endoflife care: A national study. Journal of the American Geriatrics Society. 2007;55(2):189-194.
5. Detering KM, Hancock AD, Reade MC, Silvester W. The impact of advance care planning on end of life care in elderly patients: randomised controlled trial. BMJ. 2010;340:c1345
6. Huynh TN, Kleerup EC, Raj PP, Wenger NS. The Opportunity Cost Of Futile Treatment In The Intensive Care Unit. Critical care medicine. 2014;42(9):1977-1982. doi:10.1097/CCM.0000000000000402.
7. Huynh, TN. Kleerup, EC., Wiley, JF., Savitsky, TD., Guse, D., Garber, BJ., Wenger, NS. The Frequency and Cost of Treatment Perceived to Be Futile in Critical Care. JAMA Intern Med.
8. NQF #1626: Patients Admitted to ICU Who Have Care Preferences Documented. National Quality Forum.
9. Khandelwal N., Kross E., Engelberg R., Coe N., Long A., Curtis J. Estimating the Effect of Palliative Care Interventions and Advance Care Planning on ICU Utilization: A Systematic Review. doi:10.1097/CCM.0000000000000852. Crit Care Med. 2015 May.
10. Rising JC, J.; Valuck, T. Building Additional Serious Illness Measures Into Medicare Programs. The Pew Charitable Trusts. 2017.
11. Walling AM, Tisnado D, Asch SM, et al. The quality of supportive cancer care in the veterans affairs health system and targets for improvement. JAMA internal medicine. 2013;173(22):2071-2079.
12. Dy SM, Lorenz KA, O'Neill SM, et al. Cancer Quality-ASSIST supportive oncology quality indicator set: feasibility, reliability, and validity testing. Cancer. 2010;116(13):3267-3275.
13. Aldridge MD, Meier DE. It is possible: quality measurement during serious illness. JAMA Intern Med. 2013;173(22):2080-2081.
14. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. JAMA. 2005;12(4):448-457
15. Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 13731378, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

16. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, Crockett SD, Gourevitch RA, Dean KM, Mehrotra A. Challenges in adapting existing clinical natural language processing systems to multiple, diverse healthcare settings. (JAMIA)
17. Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". Neural Networks. 61: 85117. arXiv:1404.7828. doi:10.1016/j.neunet.2014.09.003. PMID 25462637
18. Johnson AE, Pollard, T. J., Shen, L., Lehman, L. W., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., Mark, R. G. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
19. Walling AM, Asch, S.M., Lorenz, K.A., Roth, C.P., Barry, T., Kahn, K.L., Wenger, N.S. The Quality of Care Provided to Hospitalized Patients at the End of Life. Archives of Internal Medicine. 2010;170(12):1057-1063.
20. Honnibal MJ, M. An Improved Non-monotonic Transition System for Dependency Parsing. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon, Portugal.
21. Dernoncourt F, Lee, J. Y., Uzuner, O., Szolovits, P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017;24(3):596-606.
22. Dernoncourt F, Lee, J.Y, Szolovits, P. . NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. . Conference on Empirical Methods on Natural Language Processing (EMNLP). 2017.
23. Efron B. Better Bootstrap Confidence Intervals. Journal of the American Statistical Association. 1987;82(397):171-185.
24. Davison AC, Hinkley, D.V. Bootstrap Methods and their Application. Cambridge University Press; 1997.
25. D'Avolio LW, Nguyen, T. M., Goryachev, S., Fiore, L. D. Automated concept-level information extraction to reduce the need for custom software and rules development. Journal of the American Medical Informatics Association. 2011;18(5):607-613.
26. Xu H, Jiang, M., Oetjens, M., Bowton, E.A., Ramirez, A.H., Jeff, J.M.,Basford, M.A., Pulley, J.M., Cowan, J.D., Wang, X., Ritchie, M.D., Masys, D.R., Roden, D.M., Crawford, D.C., Denny, J.C. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. Journal of the American Medical Informatics Association. 2011;18(4):387-391.
27. Bejnordi BE, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Journal of the American Medical Association. 2017;318(22):2199-2210.
28. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Journal of the American Medical Association. 2016;316(22):2402-2410.
29. Ting DSW, Cheung, C.Y., Lim, G.,Tan, G.S.W., Quang, N.D., Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. Journal of the American Medical Association.
30. Charlotta Lindvall, Elizabeth J. Lilley, Sophia N. Zupanc, Isabel Chien, Alexander W. Forsyth, Anne Walling, Zara Cooper, and James A. Tulsky. Natural language processing to assess palliative care processes in cancer patients receiving palliative surgery. In preparation.

A Regular expression library

Domain	Keywords
Patient care preferences	goc, goals of care, goals for care, goals of treatment, goals for treatment, treatment goals, family meeting, family discussion, family discussions, patient goals, dnr, dni, dnrdni, dnr/dni, DNI/R, do not resuscitate, do-not-resuscitate, do not intubate, do-not-intubate, chest compressions, no defibrillation, no endotracheal intubation, no mechanical intubation, shocks, cmo, comfort measures
Goals of care conversations	goc, goals of care, goals for care, goals of treatment, goals for treatment, treatment goals, family meeting, family discussion, family discussions, patient goals
Code status limitations	dnr, dni, dnrdni, dnrdni, DNIR, do not resuscitate, do-not-resuscitate, do not intubate, do-not-intubate, chest compressions, no defibrillation, no endotracheal intubation, no mechanical intubation, shocks, cmo, comfort measures
Communication with family	Explicit conversations held during ICU stay period with patients or family members about the patients goals, values, or priorities for treatment and outcomes.
Full code status	full code

B Token-level performance

Table 4. Performance (%) of the neural network on the validation data set at the token-level.

Domain	F1-score	Accuracy	Sensitivity	Positive Predictive Value	Specificity
Patient care preferences	76.0	99.6	75.8	75.2	99.8
Goals of care conversations	70.4	99.6	70.0	69.9	99.8
Code status limitations	76.3	99.8	72.7	80.5	99.9
Communication with family	68.2	99.7	62.0	76.4	99.9
Full code status	90.9	99.8	88.3	93.6	99.8

C Examples of identified text

Below are examples of correctly identified serious illness documentation by the neural network and regular expression methods in the validation dataset. Correctly identified tokens are bolded. Typographical errors are from the original text. Each cell includes an example of identified tokens in the same text and an example of documentation identified by the neural network that was missed by the regular expression method, if relevant.

Domain	Neural Network	Regular Expression
Goals of care conversations	Hypercarbic resp failure: family meeting was held with son/**HCP and in keeping with patients goals of care, there was no plan for intubation**.Family was brought in and we explained the graveness of her ABG and her worsened mental status which had failed to improve with BiPAP. **Family was comfortable with removing Bipap and providing comfort care** including morphine prn. **family open to cmo but pt wants full code but also doesn't want treatment or to be disturbed**.	Hypercarbic resp failure: **family meeting** was held with son/HCP and in keeping with patients **goals of care**, there was no plan for intubation.Family was brought in and we explained the graveness of her ABG and her worsened mental status which had failed to improve with BiPAP. Family was comfortable with removing Bipap and providing comfort care including morphine prn. family open to cmo but pt wants full code but also doesn't want treatment or to be disturbed.
Code status limitations	**CODE: DNR/DNI, confirmed with healthcare manager who will be discussing with official HCP**	CODE: **DNR/DNI**, confirmed with healthcare manager who will be discussing with official HCP

Communication with family	Dr. [**First Name (STitle) **] from **neurosurgery held family meeting and explained grave prognosis to the family.** **lengthy discussion with the son who is health care proxy he wishes to pursue comfort measures due to severe and unrevascularizable cad daughter is not in agreement at this time but is not the proxy due to underlying psychiatric illness**	Dr. [**First Name (STitle) **] from neurosurgery held **family meeting** and explained grave prognosis to the family. lengthy discussion with the son who is health care proxy he wishes to pursue comfort measures due to severe and unrevascularizable cad daughter is not in agreement at this time but is not the proxy due to underlying psychiatric illness
Full code status	**Code: FULL; Discussed with daughter and HCP who says that patient is in a Hospice program with a "bridge" to DNR/DNI/CMO, but despite multiple conversations, the patient insists on being full code** **CODE: Presumed full**	Code: FULL; Discussed with daughter and HCP who says that patient is in a Hospice program with a "bridge" to DNR/DNI/CMO, but despite multiple conversations, the patient insists on being **full code** CODE: Presumed full