Expanding the Reach of Structured EHR Data with Clinical Notes: Improving End-of-Life Prediction

Seda Bilaloglu, Vincent J. Major, Himanshu Grover, Isabel Metzger, & Yindalon Aphinyanaphongs
NYU Langone Health, Department of Population Health
227 East 30th Street, Sixth floor, New York, NY 10016

Abstract

Appropriate treatment decisions and end-of-life planning for patients with serious, life-limiting disease rely on accurate prognosis estimates. Many existing methods rely on structured electronic health record data which may limit generalizability across sites and restrict performance for patients with less documented history. Clinical notes may help to 'level the playing field'. We use History and Physical (H&P) notes written within 16 hours of hospitalization to predict 60-day, all-cause mortality. We test several neural network approaches and observe little improvement over a CNN by adding bi-directional recurrence or convolutional attention. The CNN was prospectively validated against an existing system using structured data. The CNN reports higher discrimination (86.0% vs. 80.6%) and average precision (31.4% vs. 17.9%). The CNN identifies fewer patients at high-risk but 91% were under-estimated by the existing method (high-risk: 80 vs. 131 with overlap of 7). Patients of both groups do die in the following months suggesting the two approaches identify different patient phenotypes which supplement one another. The CNN model better adapts to a new hospital where many patients have little or no structured history incapacitating the existing system (high-risk: 27 vs. 1).

As patients with life-limiting disease approach the end of their life, physicians incorporate the patient's preferences and estimated prognosis to adapt their treatment. Patients with high symptom burden, e.g. nausea, fatigue, or pain, often elect for treatments with supportive intent and forego curative interventions, e.g. surgery or chemotherapy. However, physicians struggle to accurately predict a patient's risk, typically being optimistic (Christakis and Lamont 2000; Glare, Eychmueller, and Virik 2003; White et al. 2016), which can lead them to defer discussing prognosis or initiating any end-of-life planning and, instead, continuing aggressive treatment.

Many patients with chronic disease or advanced cancer are hospitalized at least once within their last year of life with worsening symptoms (Canadian Institute for Health Information 2011; Schifeling and Fischer 2020). While acutely ill and in the hospital, these patients are treated by clinicians unfamiliar with their disease, history, or preferences. In lieu of any explicit documentation such as advance care planning, code status or advance directives, these patients may receive unwanted aggressive treatment. Precise identification of high-risk patients can break this cycle by encouraging appropriate end-of-life care.

Clinical risk tools often provide a score (Charlson et al. 1987; Knaus et al. 1985; Morita et al. 1999) to stratify patients into risk groups. Numerous machine learning methods also exist but many are limited to specific populations by disease or acuity (Ghassemi et al. 2014; Makar et al. 2015; Elfiky et al. 2017; Parikh et al. 2019). Several general approaches have been proposed for use to prompt clinicians to consider end-of-life planning (Avati et al. 2018; Wegier et al. 2019; Courtright et al. 2019; Major and Aphinyanaphongs 2020). Each of these works rely on structured electronic health record (EHR) data to generate their predictions. The reliance on EHR data is expected to under-serve populations with less access to care and, thus, less structured data.

Objective

The objective of this work is to investigate how a text-based approach compares to one using coded structured EHR data. A predictive model using a single History & Physical examination (H&P) note is developed to estimate risk of death within two months. H&P notes are available for all hospitalized patients shortly after admission. A variety of neural network architectures and experimental settings will be tested before implementing one model to make daily predictions on newly admitted patients. This model will be compared to an existing system deployed at our institution based on an approach using coded structured EHR data available at the time of admission (Major and Aphinyanaphongs 2020).

Related works

Recent advances in neural networks have already improved how natural language processing systems can understand and generate human language. Convolutional neural networks (CNNs) (Lecun et al. 1998) encode spatially invariant features and have shown success in sentence classification (Kim 2014). Further advances such as Convolutional Attention for Multi-Label classification (CAML) (Mullenbach et al. 2018) extend how the network can learn to focus on a subset of the text. Recurrent neural networks (RNNs) naturally represent the sequential nature of text and excel in text classification (Yin et al. 2017).

Recent studies have shown neural networks and clinical text may be useful for applications such as diagnosis code assignment (Mullenbach et al. 2018) and predicting emergency department disposition (Sterling et al. 2019). Other work has reported improvements in performance when using clinical notes to predict sepsis (Culliton et al. 2017) compared to structured data where combining both modalities may be best. However, not all clinical notes are equal. For example, some notes such as triage notes are brief and focus on a patient's chief complaint and vitals rather than their clinical history. Related work to predict mortality overcomes some limitations of short, heterogeneous clinical notes by combining learned topics from several or many notes from a patient's history (Ghassemi et al. 2014; Wang et al. 2019).

Methods

Cohort and experimental design

The dataset for this work is a subset of clinical notes from a large US academic medical center spanning three hospitals over five years, January 1st, 2013 to December 31st, 2017. In this time, 87,293 unique adult patients were hospitalized 128,328 times. Death outcomes include institutional and Social Security deaths as well as initiation of hospice. Positive cases are patients who have a documented date of death within 60 days of admission, while negative cases are those who do not (i.e. no censoring requirements were imposed). Separation into training, validation, and testing sets was performed temporally and at the patient level to prevent potential for data leakage between sets (Neto et al. 2019).

H&P notes

An H&P note is the result of a review of the patient's medical history and a comprehensive physical examination conducted by the authoring physician. A typical H&P note begins with a subjective section that describes the patient's chief complaint, symptoms and history followed by an objective section that includes vitals, physical exam findings, labs, and imaging. The author typically concludes with a narrative describing their clinical assessment and plan for treatment or testing.

H&P notes are typically written within the first day of a hospital admission by a relatively senior member of the care team. We restrict to the most common author types (which together constitute 98%), namely: Physician, Fellow, Resident, Physician Assistant, and Nurse Practitioner. Multiple H&P notes may be written during one hospitalization, e.g. by different authors or departments. For this work we restricted to H&P notes created between 0–16 hours of admission and those with more than 50 words (removing many addendum and attestations; typical H&P note has 1000–2000 words).

Text preprocessing Some sections of H&P notes are heterogeneous, varying widely by the patient's history and presentation, other sections are consistently present and similarly described between authors. One section, History of Present Illness, is extracted from the larger note by a set of simple—but effective—rules (Figure 1). The remaining text is preprocessed by standard techniques to remove punctuation, trim whitespace, and mask numbers. For example: "History of Present Illness: This is a 58 y.o. female current smoker, HLD with c/o occasional dizziness while lying down and turning quickly. Denies any syncope, LOC, CVA, vision changes." Becomes: "history of present illness this is a _num_ y o female current smoker hld with c o occasional dizziness while lying down and turning quickly denies any syncope loc cva vision changes"

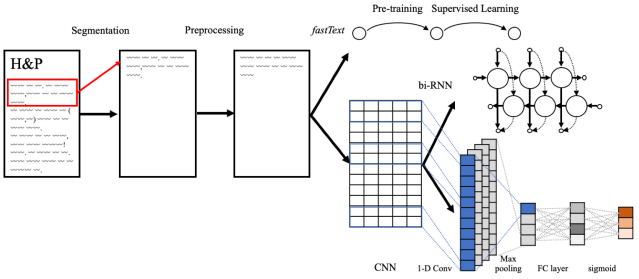

Figure 1: Preprocessing and model development workflow

Baseline linear classifier

We employed *fastText* (Bojanowski et al. 2016) to pre-train word embeddings with a large, unlabeled corpus of medical text similar to related work (Xu et al. 2018). The corpus included all training set H&P notes (62k documents and 50M words) and all clinical notes within the MIMIC-III Critical Care Database (Johnson et al. 2016) (2M documents and 5M words). The MIMIC-III notes are expected to be written in a similar way to the H&P notes with similar language of domain-specific terms and acronyms including context related to mortality risk. *fastText* skip-gram embeddings were tested with various settings (optimization of negative sampling vs. hierarchical softmax and 300 vs. 600 dimensions) but otherwise default parameters (e.g. max_n=8, epoch=20). Pre-trained embeddings were used to learn a *fastText* linear classifier (Joulin et al. 2016) for 60 day mortality testing parameters of vector dimension, number of epochs, and optimization method.

Neural network architectures

We tested various neural network architectures, two CNN variants (with and without CAML (Mullenbach et al. 2018)), and one RNN (testing both bi-directional LSTM (Hochreiter and Schmidhuber 1997) and GRU (Cho et al. 2014) variants). Similar to related work (Mullenbach et al. 2018), we tested various hyperparameters with grid search comparing validation set performance. We tested CNN hyperparameters of: embedding dim={100, 300}, number of kernels={50, 100, 200}, kernel size={5, 10}, activation function={relu, tanh}, and lr={1e-3, 3e-3}. And tested RNN hyperparameters of: recurrence unit={LSTM, GRU}, embedding dim={100, 300}, hidden layer length={128, 256}, number of layers={1, 2}, lr={1e-3, 3e-3}.

All architectures were implemented using PyTorch (Paszke et al. 2017), employed a binary cross entropy loss, used the Adam optimizer (Kingma and Ba 2014), with early stopping enabled (no improvement for three consecutive epochs). Moreover, each model used: vocabulary size=20,000, input length=2,000, and max epochs=200. The CNNs used max pooling, one fully connected layer, batch size=512, with dropout=0.2 and the RNNs used batch size=64, and activation function=*tanh*.

Evaluation metrics

End-of-life is a rare outcome which can skew evaluation metrics such as accuracy. Instead, we use discrimination, visualized with receiver operating characteristic (ROC) and measured by the area under ROC (AUROC). As this model is potentially helpful to recommend an intervention to predicted positives, we also employ precision-recall curves (PRC) and measure the area under PRC (AUPRC) and the max-F1 score.

Prospective validation

One model would be validated prospectively by generating predictions daily for recently admitted patients. To ease comparison with an existing system that uses structured data, an operating threshold was selected for 60-day mortality at 75% precision. Over an equal time period, the two models were compared in terms of their discrimination, average precision, percentage of admitted patients who receive any prediction, the number of patients identified above the high-risk threshold, and the subsequent survival of those high-risk patients.

Results

Cohort

The final cohort (Table 1) included a total of 82,788 unique hospitalizations of adult patients within at least one sufficiently long H&P note ($\geq$ 50 words) created within 16 hours of admission. The training set contains four years of data with the validation and test sets six months each. Mortality within 60 days of admission is observed in 5.1% of cases.

fastText linear classifier

We found *fastText* AUROC was insensitive to tested settings. The (marginally) best performing model reported test set performance of AUROC of 0.903, AUPRC of 0.347, and max F-1 of 0.379 using hierarchical sampling, embeddings of dimension 600, and 50 training epochs. The pre-trained embeddings added little over random initialization, suggesting the dataset of H&P notes contains sufficient information.

Neural network architectures

Overall, increased CNN performance was observed with larger embedding dimensions (300), more kernels (200), smaller kernel size (5), larger learning rate (3e-3) and ReLU activation (vs. tanh). Improved RNN performance was observed with GRU (vs. LSTM), larger embedding dimension (300), larger hidden layer length (256), smaller number of layers (1) and larger learning rate (3e-3). Overall, the ROCs and PRCs were similar for different architectures (Figure 2 and Table 2). When we compared the speeds, we found that CNN was 40–75% faster than CAML and RNN. Moreover, when bootstrapping for an operating threshold at 75% precision, there was no notable improvement in recall by the RNN or CAML over the CNN. Therefore, we chose to prospectively validate the simpler, quicker CNN model.

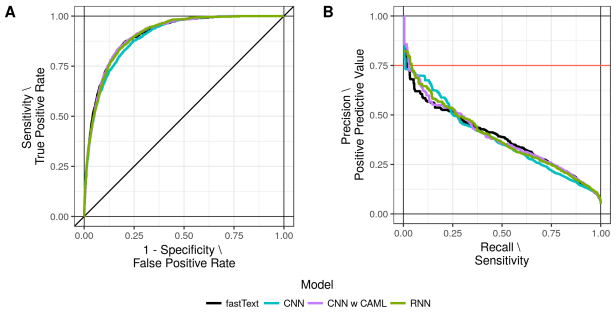

Figure 2: Test set evaluation metrics. A) ROC and B) PRC.

Prospective validation

The CNN model was validated *prospectively* by predicting risk for newly hospitalized patients at four hospitals of the same institution (one additional than development). The same note and patient eligibility criteria as model development (i.e. age $>$ 18 years, H&P notes created within 16 hours of admission, $\geq$ 50 words, etc.) were applied to new H&P notes each morning for nine months. The structured data approach was similarly applied over this period.

During these nine months, a total of 65,727 eligible patients were hospitalized and while the CNN model made more predictions than the structured data model (57,997 vs. 53,446; Table 3), fewer unique patients were scored (37,720 vs. 53,446). The CNN makes predictions for thousands of patients who are missed by the structured data approach due to insufficient data (9,344, 14.2% of all admissions; Figure 3A). The CNN outperformed the structured model in discrimination and average precision (Table 3). The CNN is delayed a median of 29.6 hours after admission to allow the H&P to be created, written, signed, and available in the database compared to minutes of the structured data approach (Table 3). The additional delay to access signed notes could not feasibly be shortened and did not effect patient eligibility or model performance (except timing).

A preselected threshold at 75% precision allows each model to identify patients as 'high-risk'. The CNN finds 80 patients, but only seven overlap with the 131 identified by the structured data approach (Figure 3B). Of the additional 73 high-risk cases added by the CNN, 11 are patients with no structured data prediction due to insufficient data (i.e. from the 9,344 of 3B) and the remaining 62 were estimated as low-risk by the structured data model.

Generalization to a new site One of the four hospitals included in the prospective validation was new to our institution's EHR system and not used to develop either model.

Table 1: Datasets of H&P notes and 60-day patient outcomes used for model development.

	Training	Validation	Test	Total	
Positive	2,679	468	1,055	4,202	(5.1%)
Negative	50,140	8,973	19,473	78,586	
Total	52,819	9,441	20,528	82,788	

Table 2: Test set evaluation metrics.

Model	**AUROC** [95% CI]	**AUPRC** [95% CI]	**max F-1** [95% CI]
CNN	0.899 [0.890, 0.908]	0.381 [0.348, 0.421]	0.418 [0.394, 0.450]
RNN (bi-GRU)	0.907 [0.899, 0.915]	0.388 [0.357, 0.427]	0.421 [0.399, 0.452]
CNN with CAML	0.908 [0.900, 0.917]	0.388 [0.354, 0.424]	0.425 [0.401, 0.454]

Table 3: Prospective validation results.

Metric	**H&P CNN**	**Structured Data**
Total admissions	65,727	
Predictions	57,997	53,446
Admissions predicted	37,720 (57.4%)	53,446 (81.3%)
Timing (hrs) median [IQR]	29.6 [19.0, 36.8]	0.03 [0.02, 0.85]
High-risk admissions	80 (0.21%)	131 (0.25%)
AUROC [95% CI]	0.860 [0.847, 0.873]	0.806 [0.793, 0.820]
AUPRC [95% CI]	0.314 [0.282, 0.353]	0.179 [0.157, 0.204]
Max F-1 [95% CI]	0.377 [0.352, 0.409]	0.228 [0.210, 0.254]

Patients hospitalized at this new site had less structured data than expected as their prior care was captured in another EHR system. This unintentional ablation hinders the structured data model: generating fewer predictions and underestimating risk when predictions could be made.

During this period, 21,101 admissions occurred at the new site where the CNN and structured data models make predictions for 53.2% (11,234) and 65.6% (13,841) of admissions. Adding the CNN to the structured data approach adds 2,336 admissions (11.1% of total) that would otherwise have no prediction made. Overall, the CNN produced better discrimination, 85.2% (95% CI: 82.6–87.5) vs. 77.3% (95% CI: 74.7–79.7). A total of 27 high-risk patients were identified at the new site where only one was found by the structured data model and the remaining 26 by the CNN.

Survival Outcomes for each identified high-risk patient were censored with at least 180 days of follow-up, reducing the period to five months (n_{CNN} = 54 and $n_{structured}$ = 62). Survival analysis (Figure 3C) finds median survival of both groups is shorter than 60 days but CNN survival was significantly shorter. Only 9.3% of the CNN group remained alive and at risk 60 days after admission compared to 32%. There was no evidence observed that communicating predictions to providers affected (either to improve or worsen) survival outcomes, but any change is expected to equally affect patients found by either approach.

Combining approaches Neither the structured data nor the H&P CNN approaches are perfect; both miss patients when data is not available while making predictions the other cannot. Combining the two approaches adds redundancy by increasing the proportion of admissions with any score and the number of patients identified as high-risk who may benefit from end-of-life planning. The two models are more akin to *siblings* than *alternatives* as they can cooperate to achieve the same goal. Emsembling the two methods (or other multi-modal approaches) are hindered by the median 30 hour delay between availability of data and would likely underserve patients with either type of missing data. By combining their outputs with OR logic, the number of patients identified jumps from 131 to 184, a 55.7% increase.

Discussion

Many AI approaches in medicine rely on complex structured data within the EHR. Reliance on this EHR data can overfit to nuances of the data created by the EHR vendor, institutional settings, or data capture mechanisms. The use of multi-site datasets may mitigate these biases but the ability of models to generalize to other sites remains understudied. Moreover, model fairness—how different patient populations are impacted—is only beginning to be studied in medicine, expanding on work from other fields (Corbett-Davies and Goel 2018; Mitchell et al. 2018). While how one model affects patients of different genders for example is critically important, so too is how unseen confounders such as access to care impact the data used to inform such predictions. The latter is not as simple as perturbing race (Obermeyer et al. 2019) and is a symptom of inequity.

Structured data approaches that rely on documented clinical history such as billing codes use these features to describe clinical history exploiting a correlation between utilization and risk of death. Patients with little structured data will receive low estimates of risk. However, patients new to the institution, those who receive their care split between two systems, or whose community has limited access to care would be unfairly disadvantaged.

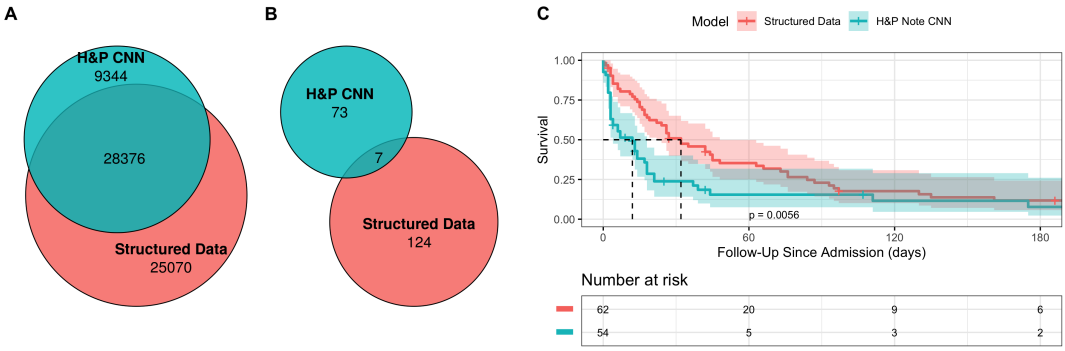

Figure 3: Comparison of prospective results of the H&P CNN and the structured data models. A) admissions with at least one prediction, B) admissions identified as high-risk, and C) survival analysis of high-risk patients with 180 days of follow-up.

The objective of this work was to identify opportunities to build on a previously implemented system with zero adaptations of existing clinical workflows. While higher performance may be possible with a system which periodically updates with additional data, this was not feasible in our setting. A multi-modal model including both structured and unstructured data was considered but exceeded our current ability to implement in real-time. Instead, a single text-based prediction as close to admission as possible was desired to elegantly supplement the existing system which alerts attending physicians of identified high-risk patients.

In this work we present one solution to bridge the inequity of structured EHR data by using no clinical history and only a single H&P note to predict near-term mortality. We find high numbers of patients newly predicted or newly identified as high-risk by the H&P model compared to a structured data approach. These patients were disadvantaged by the structured data model. Prospective validation finds higher discrimination and average precision by the H&P CNN (Table 3) but both approaches find patients who do die within 60 days of admission (Figure 3C). In fact, those found by the H&P model have significantly shorter survival suggesting the selected operating threshold could be relaxed to identify even more high-risk patients.

In a natural experiment enabled by a new hospital transitioning into the existing EHR system, the H&P approach generalizes much better to patients with effectively ablated clinical histories. Use of the preselected threshold identifies 26 high-risk patients using the H&P approach vs 1.

Comparing text approaches

The CNN model prospectively validated was the simpler, quicker neural network tested; little improvement was observed with convolution attention or recurrence (Table 2). Recent works have reported that RNNs outperform CNNs in various text classification tasks (Yin et al. 2017), including predicting the onset of various diseases (Liu, Zhang, and Razavian 2018), and CAML improves performance over both CNN and RNNs (Mullenbach et al. 2018). CNN and RNN architectures approach text classification in different ways. Whereas CNNs can identify localized signals from specific keywords or phrases, RNNs flexibly model context dependencies. Their differences force the two architectures to learn in different ways, despite likely focusing on similar phrases. For mortality prediction, some keywords are undoubtedly associated with high-risk patients, e.g. “terminal”, “hospice”, or “palliative”. For this task, the differences between CNNs and RNNs have little effect on performance.

A baseline *fastText* linear classifier, produced a similar AUROC but lower AUPRC and max F-1 as a result of poor performance in the high precision, low recall area of the PRC. Upon bootstrapping, a trivial 75% precision threshold at 0% recall was frequently selected—rendering the model useless in practice. The added complexity of the CNN enables the model to learn patterns beyond the linear sum of embeddings used by *fastText* to correctly rank very high-risk patients without false positives.

Limitations

This work is one specific application and should not be used to discredit the use of billing or structured EHR data. Clinical notes are not immune from bias which may lead to inequitable predictions. Well documented issues of mistrust (Boag et al. 2018) and undertreatment (Parikh et al. 2020) by race or ethnicity are likely embedded in H&P notes. Future work will explore if the CNN model has learnt these sources of bias and quantify the relative fairness of the two alternative approaches.

Performance of the text-based models may have been affected by our experimental choices. The segmentation of the History of Present Illness section of the notes omits other information which may be useful for the RNN for example. Secondly, our preprocessing tokenizes numbers and strips symbols which may carry useful, albeit highly specific, information (Cruz Díaz and Maña López 2015) such as medication or radiation dosage. Thirdly, experimentation did not include simple keyword approaches (as initial testing found poor results) or recent methods such as transformer networks and BERT (as these methods were not widely used at the time).

Conclusion

We show that a hospitalized patient's H&P note is a feasible means to identify patients at risk of dying and can be used to prompt end-of-life planning. Moreover, the text-based approach outperforms an existing system in prospective validation and generalizes better to a new hospital location. Adding the text-based approach to our current system improves the system's reach by scoring and identifying more patients while also adding desirable redundancy.

References

Avati, A.; Jung, K.; Harman, S.; Downing, L.; Ng, A.; and Shah, N. H. 2018. Improving palliative care with deep learning. *BMC Med. Inform. Decis. Mak.* 18(Suppl 4):122.

Boag, W.; Suresh, H.; Celi, L. A.; Szolovits, P.; and Ghassemi, M. 2018. Racial disparities and mistrust in End-of-Life care.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information.

Canadian Institute for Health Information. 2011. Healthcare use at the end of life in atlantic canada.

Charlson, M. E.; Pompei, P.; Ales, K. L.; and MacKenzie, C. R. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* 40(5):373–383.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation.

Christakis, N. A., and Lamont, E. B. 2000. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ* 320(7233):469–473.

Corbett-Davies, S., and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning.

Courtright, K. R.; Chivers, C.; Becker, M.; Regli, S. H.; Pepper, L. C.; Draugelis, M. E.; and O'Connor, N. R. 2019. Electronic health record mortality prediction model for targeted palliative care among hospitalized medical patients: a pilot quasi-experimental study. *J. Gen. Intern. Med.* 34(9):1841–1847.

Cruz Díaz, N. P., and Maña López, M. J. 2015. An analysis of biomedical tokenization: Problems and strategies. *undefined*.

Culliton, P.; Levinson, M.; Ehresman, A.; Wherry, J.; Steingrub, J. S.; and Gallant, S. I. 2017. Predicting severe sepsis using text from the electronic health record.

Elfiky, A.; Pany, M.; Parikh, R.; and Obermeyer, Z. 2017. A machine learning approach to predicting short-term mortality risk in patients starting chemotherapy.

Ghassemi, M.; Naumann, T.; Doshi-Velez, F.; Brimmer, N.; Joshi, R.; Rumshisky, A.; and Szolovits, P. 2014. Unfolding physiological state: Mortality modelling in intensive care units. *KDD* 2014:75–84.

Glare, P.; Eychmueller, S.; and Virik, K. 2003. The use of the palliative prognostic score in patients with diagnoses other than cancer. *J. Pain Symptom Manage.* 26(4):883–885.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-W. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Sci Data* 3:160035.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification.

Kim, Y. 2014. Convolutional neural networks for sentence classification.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization.

Knaus, W. A.; Draper, E. A.; Wagner, D. P.; and Zimmerman, J. E. 1985. APACHE II: a severity of disease classification system. *Crit. Care Med.* 13(10):818–829.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11):2278–2324.

Liu, J.; Zhang, Z.; and Razavian, N. 2018. Deep EHR: Chronic disease prediction using medical notes.

Major, V. J., and Aphinyanaphongs, Y. 2020. Development, implementation, and prospective validation of a model to predict 60-day end-of-life in hospitalized adults upon admission at three sites. *BMC Med. Inform. Decis. Mak.* 20(1).

Makar, M.; Ghassemi, M.; Cutler, D. M.; and Obermeyer, Z. 2015. Short-term mortality prediction for elderly patients using medicare claims data. *Int J Mach Learn Comput* 5(3):192–197.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2018. Prediction-Based decisions and fairness: A catalogue of choices, assumptions, and definitions.

Morita, T.; Tsunoda, J.; Inoue, S.; and Chihara, S. 1999. The palliative prognostic index: a scoring system for survival prediction of terminally ill cancer patients. *Support. Care Cancer* 7(3):128–133.

Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; and Eisenstein, J. 2018. Explainable prediction of medical codes from clinical text.

Neto, E. C.; Pratap, A.; Perumal, T. M.; Tummalacherla, M.; Snyder, P.; Bot, B. M.; Trister, A. D.; Friend, S. H.; Mangravite, L.; and Omberg, L. 2019. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *npj Digital Medicine* 2(1):1–6.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453.

Parikh, R. B.; Manz, C.; Chivers, C.; Regli, S. H.; Braun, J.; Draugelis, M. E.; Schuchter, L. M.; Shulman, L. N.; Navathe, A. S.; Patel, M. S.; and O'Connor, N. R. 2019. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Netw Open* 2(10):e1915997.

Parikh, R. B.; Robinson, K. W.; Chhatre, S.; Medvedeva, E.; Cashy, J. P.; Veera, S.; Bauml, J. M.; Fojo, T.; Navathe, A. S.; Bruce Malkowicz, S.; Mamtani, R.; and Jayadevappa, R. 2020. Comparison by race of conservative management for Low-Risk and Intermediate-Risk prostate cancers in veterans from 2004 to 2018.

Paszke, A.; Gross, S.; Chintala, S.; and Chanan, G. 2017. Pytorch. *Computer software. Vers. 0. 3* 1.

Schifeling, C. H., and Fischer, S. M. 2020. Missing the mark: High rates of absent and untimely access to specialty palliative care in patients with Peri-Hospital mortality. *J. Palliat. Med.*

Sterling, N. W.; Patzer, R. E.; Di, M.; and Schrager, J. D. 2019. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int. J. Med. Inform.* 129:184–188.

Wang, L.; Sha, L.; Lakin, J. R.; Bynum, J.; Bates, D. W.; Hong, P.; and Zhou, L. 2019. Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions.

Wegier, P.; Koo, E.; Ansari, S.; Kobewka, D.; O'Connor, E.; Wu, P.; Steinberg, L.; Bell, C.; Walton, T.; van Walraven, C.; Embuldeniya, G.; Costello, J.; and Downar, J. 2019. mHOMR: a feasibility study of an automated system for identifying inpatients having an elevated risk of 1-year mortality. *BMJ Qual. Saf.*

White, N.; Reid, F.; Harris, A.; Harries, P.; and Stone, P. 2016. A systematic review of predictions of survival in palliative care: How accurate are clinicians and who are the experts? *PLoS One* 11(8):e0161407.

Xu, H.; Wang, W.; Liu, W.; and Carin, L. 2018. Distilled wasserstein learning for word embedding and topic modeling. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc. 1716–1725.

Yin, W.; Kann, K.; Yu, M.; and Schütze, H. 2017. Comparative study of CNN and RNN for natural language processing.