Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation

Kexin Huang[*1], **Abhishek Singh**[*2], **Sitong Chen**[*1], **Edward T. Moseley**[3],
Chih-Ying Deng[3], **Naomi George**[4], **Charlotta Lindvall**[34]

[1]Harvard, [2]MIT, [3]Dana Farber Cancer Institute, [4]Brigham and Women's Hospital

kexinhuang@hsph.harvard.edu, abhi24@mit.edu,
sitong_chen@hms.harvard.edu, nrgeorge@bwh.harvard.edu,
{edward_moseley,chih-ying_deng}@dfci.harvard.edu,
charlotta_lindvall@dfci.harvard.edu

Abstract

Clinical notes contain rich information, which is relatively unexploited in predictive modeling compared to structured data. In this work, we developed a new clinical text representation Clinical XLNet that leverages the temporal information of the sequence of the notes. We evaluated our models on prolonged mechanical ventilation prediction problem and our experiments demonstrated that Clinical XLNet outperforms the best baselines consistently. The models and scripts are made publicly available.

1 Introduction

Unstructured clinical notes within Electronic Health Records (EHR) contain valuable information to support clinical decisions (Murdoch and Detsky, 2013). However, most prognostic models used in medical practice currently rely on scoring systems that only incorporate structured data (Gall et al., 1986; Gall, 1993; Vincent et al., 1996; Rapsang and Shyam, 2014).

A major challenge to utilize unstructured clinical data is in representing notes in ways that allow effective mining of clinically meaningful knowledge. There are many recent advances in the standard Natural Language Processing domain, such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019). However, clinical notes are far different from the general domain text (Wikipedia, BookCorpus, etc). For example, clinical notes contain jargon and abbreviations, different grammar and syntax. It is notoriously difficult to obtain an effective note representation. Recently, ClinicalBERT, which adapts the BERT model from the standard NLP domain to model clinical notes (Huang et al., 2019; Alsentzer et al., 2019) achieved superior performance in clinical text prediction. However, previous works still have the following limitations:

- **Notes representation could be improved.** In the standard NLP domain, BERT ignores the discrepancy of masked positions between the pretraining and finetuning stage. An autoregressive pretraining method named XLNet was recently developed and empirically outperforms BERT on many NLP tasks (Yang et al., 2019).

- **Failure to incorporate the temporal dimension of clinical notes.** Clinical notes have a temporal dimension where the order of information in sequential notes can provide additional predictive signals. Most previous models (Huang et al., 2019; Alsentzer et al., 2019) only aggregate individual risk scores from each note which ignore the temporal information charted in EHR.

In this paper, we present Clinical XLNet, which processes a patient's notes and predicts clinical outcomes. In particular, this model mitigates the aforementioned limitations via the following technical contributions:

- **Improved clinical notes representation.** We apply the permutation language modeling method proposed in XLNet on a corpus of clinical notes to generate better clinical embeddings, as demonstrated in Section. 4.

- **Inclusion of temporal information.** We maintain the temporal order of the note embeddings generated from Clinical XLNet and feed them into a bidirectional LSTM layer (Hochreiter and Schmidhuber, 1997), which leverages information along the temporal dimension (Fig. 1).

We examined Clinical XLNet's performance on a new but important clinical NLP task: predicting

Table 1: Cohort Statistics. For continuous variable, it reports mean with the standard deviation. For categorical variable, the count is given with percentage.

Statistics	All	MV $\geq$ 7d	MV < 7d	Survive < 90d	Survive $\geq$ 90d
Stays/Admissions	7,287	3,412	3,875	2,680	4,607
Age	64.3 (16.7)	63.8 (16.6)	64.8 (16.9)	69.3 (15.4)	61.4 (16.8)
Male	4,072 (55.8)	1,936 (56.7)	2,136 (55.1)	1,474 (55.0)	2598 (56.4)
Ethnicity					
White	5,159 (70.8)	2,431 (71.2)	2,728 (70.4)	1,844 (68.8)	3,315 (72.0)
Black	590 (8.1)	259 (7.6)	331 (8.5)	202 (7.5)	388 (8.4)
Hispanic/Latino	215 (2.9)	89 (2.6)	126 (3.3)	51 (1.9)	164 (3.6)
Asian	150 (2.1)	65 (1.9)	85 (2.2)	54 (2.0)	96 (2.1)
Others	1,173 (16.1)	568 (16.6)	902 (23.3)	529 (19.7)	644 (14.0)
Notes					
Word Count	1774 (1645)	1745 (1610)	1799 (1674)	1811 (1730)	1753 (1593)
Note Count	9.78 (4.70)	9.54 (4.51)	10.0 (4.86)	9.72 (4.70)	9.82 (4.70)

prolonged mechanical ventilation (PMV). Mechanical ventilation is to use an artificial breathing machine to assist or replace patient's breath. PMV stands for longer than normal period of mechanical ventilation. PMV consumes substantial healthcare resources, results in financial and emotional burdens for patients and their families, and is associated with high one-year mortality around 50-60% (Mcgee, 2010; Nelson et al., 2015; Unroe, 2010). It is projected that over 600,000 patients in the United States will require PMV by 2020 (Zilberberg et al., 2008).

Patients with PMV can have a surgical procedure called tracheostomy to establish better airway access for long-term mechanical ventilation (Cox et al., 2004). Tracheostomy is associated with improved patient comfort, decreased duration of ICU and hospital stay, and reduced mortality (Mallick and Bodenham, 2010). However, the decision to place a tracheostomy is challenging for two main reasons: (1) tracheostomy may not be necessary if a patient's condition improves without requiring PMV, and (2) tracheostomy may not be helpful if the patient is at high risk of short-term mortality. Thus, an *early* and *correct* decision of tracheostomy is critical and depends on the likelihood of PMV and short-term mortality. Accurately predicting these factors using patients' clinical notes could support clinical decision making.

We compared Clinical XLNet with several state-of-the-art baselines including BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and Clinical-BERT (Huang et al., 2019) on both PMV and mortality predictions. We performed meticulous cohort curation in MIMIC-III dataset (Johnson et al., 2016) to set up an actionable prediction task ac-

cording to the real clinical setting. Clinical notes used in this prediction task were strictly within the 48-hour time window starting from the initial mechanical ventilation event. Experimental results showed that Clinical XLNet outperformed the best baselines consistently (Section. 4).

2 Data

We use the Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC-III) dataset (Johnson et al., 2016) hosted on PhysioNet (Goldberger et al., 2000) for our model development and experiment. It consists of 61,532 ICU stays out of 58,976 hospital admissions from 46,476 patients in the intensive care unit of the Beth Israel Deaconess Medical Center (BIDMC) between 2001 and 2012 and it has 2,083,180 clinical note events.

Cohort Selection. Comprehensive inclusion and exclusion criteria were applied to the MIMIC-III dataset to generate our patient cohort who were above 18 years old, and were on mechanical ventilation for at least 2 days with more than 6 hours each day. We excluded patients who were organ donors or transferred patients from other hospitals. As certain diseases always lead to PMV, to alleviate confounding, we further removed patients with neuromuscular disease, head and neck cancer, and extensive burns (Oakden-Rayner et al., 2020). For each hospital admission, we used the first ICU stay. For clinical notes, we included nursing and respiratory notes within 48 hours from the start of the first ventilation event. The reason for only selecting nursing-related notes was to expand the cohort as MIMIC-III is missing physician notes from 2001 to 2008. Additional criteria that are applied in our data curation process are in Fig. 2 in the appendix.

In the end, we obtained a cohort of 7,287 unique patients and their corresponding 73,224 clinical notes. Table. 1 shows the cohort demographics.

Cohort Labels. Our cohort was labeled with PMV and 90-day mortality as a binary outcome. PMV was defined as being on mechanical ventilation for more than 7 days with at least 6 hours each day (Boles JM, 2007). Short-term mortality was defined as death occurring within 90 days of the first ICU admission. We use 90 days in contrast to 30 days because recent studies showed that 30 days mortality may underestimate the evaluation (Mise et al., 2015; Hirji et al., 2020).

3 Methods

This section presents our Clinical XLNet framework (Fig. 1). Clinical XLNet is an extension of XLNet (Yang et al., 2019) on the clinical text domain. It first generates a deep latent representation for clinical notes and then by applying a bidirectional LSTM (Bi-LSTM) layer, it also leverages the sequential order of notes.

Problem Settings. Our target task aims to leverage a patient's clinical notes to predict variables such as PMV and mortality. We denote a patient as P, and each patient P is associated with an ordered sequence of notes $\{N_1, \cdots, N_i\}$, where i is the total number of notes.

To predict mortality L_M and PMV L_P, we aim to learn two mappings $\mathcal{M} : \{N_1, \cdots, N_i\} \longrightarrow [0, 1]$ and $\mathcal{P} : \{N_1, \cdots, N_i\} \longrightarrow [0, 1]$ where $[0, 1]$ is a probability that measures the likelihood of having mortality and PMV respectively.

Pretraining Clinical XLNet. The XLNet is pretrained on common language corpora such as BookCorpus, Wikipedia, Common Crawl and etc. However, these corpora are different from clinical notes which are filled with jargon, abbreviations and difficult syntax and grammar. Hence, to learn an effective representation of clinical notes, we further pre-trained the XLNet using nursing, nursing/others, and respiratory therapy notes available in the MIMIC-III dataset. The clinical notes used in pre-training were NOT in the holdout test set to avoid biased results.

XLNet is a stack of Transformer-XL encoder (Dai et al., 2019). For pre-training, it uses Permutation Language Modeling (PLM) to tackle the challenge of [MASK] token information gap

between pre-training and finetuning in BERT (Devlin et al., 2019). For each sequence, it is appended with a [CLS] classification token at the beginning of the sequence for downstream task usage. For a more detailed description, we refer the readers to the original paper (Yang et al., 2019).

A patient is associated with many notes, and each notes length varies. Since XLNet can only take 512 maximum tokens, we follow (Huang et al., 2019) to first concatenate all the notes and partition them into snippets N_i. Then, we use the last encoder layer hidden representation E_i of the [CLS] token to represent the note snippet. As we train with a supervised signal in the downstream task, [CLS] token would gather useful information in the entire note sequence due to the Transformer-XL's self-attention mechanism. Now, given a temporally ordered sequence of note snippets associated with a patient, we obtain a temporally-ordered sequence of notes representations $\{E_1, \cdots, E_i\}$.

Finetuning Clinical XLNet. To leverage the temporal information among the note snippets, we feed $\{E_1, \cdots, E_i\}$ into a sequential modeling layer. Specifically, we use Bi-LSTM model (Graves and Schmidhuber, 2005; Hochreiter and Schmidhuber, 1997). We use bidirectional model because not only the latter notes depend on the previous notes as patients develop their symptoms in a temporal order but also the latter notes may contain useful clinical knowledge to help enrich the representation of previous notes. The output of the Bi-LSTM layer H_N is then fed into a predictor neural network, which at last, generates a probability p that measures the likelihood of downstream target variable, PMV, and mortality. The network is then tuned using binary classification loss.

Pre-Finetuning. As each patient is associated with many notes snippets and each snippet corresponds to a large model, it is computationally infeasible to train the model end-to-end. In order to alleviate the memory cost of end-to-end training for every clinical note, we propose to approximate the task-specific note representation through an additional pre-finetuning stage. We assume that any part of the notes associated with a patient is correlated to the label. Thus, during pre-finetuning, we use one piece of note as input, and further train the pre-trained Clinical XLNet through the downstream task label signal from the corresponding patient. The pre-finetuned network can then generate
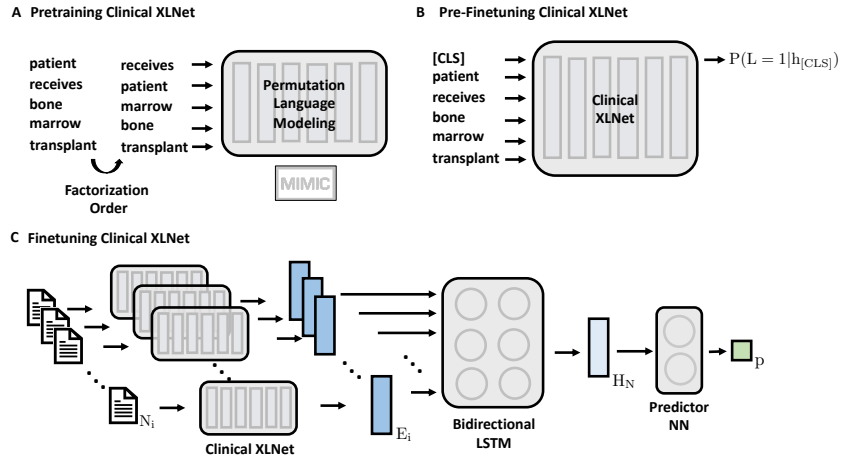
Figure 1: Clinical XLNet framework. A. We first pre-trained the XLNet embedding with MIMIC-III clinical notes dataset using Permutation Language Modeling. After pre-training, given a clinical note, the model outputs a numerical vector to be used as a note representation. B. To alleviate the computation burden from training end-to-end, the pre-finetuning stage uses the supervised signal to further tune the pre-trained network with input consisting of individual note N_i. The pre finetuned stage then generates a static task-specific note representation. C. Given a sequence of a patient's notes $\{N_1, \cdots, N_i\}$, the pre-finetuned Clinical XLNet network generates a sequence of representation of notes $\{E_1, \cdots, E_i\}$. The ordered representation sequence is then fed into a bidirectional LSTM layer, which then outputs a fixed size latent vector H_N, representing the entire sequence. H_N is finally fed into a predictor neural network to generate a probability p measuring the likelihood of the target variable.

a task specific note representation. Then, during the finetuning stage, we freeze the note representation module, and use a static fixed note representation from the pre-finetuned Clinical XLNet to feed into the Bi-LSTM layer.

4 Experiments

To evaluate our model, we examined the prediction performance under a realistic setup. We used the first 48 hours of clinical notes starting from the initial mechanical ventilation event to predict two variables: mechanical ventilation longer than 7 days and 90-days mortality. [1]

Hyperparameters. For the data split, we first obtain a 10% holdout test set. Then we generate different 8:1 train:validation splits using different random seeds for model performance robustness examination. For the pre-training, we further pre-train the XLNet embedding for another 200K steps using 16 batch size. For the pre-finetuning, we use 32 batch size with learning rate 1e-5 for four epochs with early stopping on the area under the receiver operating characteristic curve (AUROC) score of validation. For finetuning, we used a two layers Bi-LSTM module with batch size 128 and learning rate 1e-4. The pre-training and pre-finetuning pro-

Table 2: AUROC result with three independent data splits mean and standard deviation.

Method	PMV	Mortality
LSTM	0.613 ± 0.006	0.590 ± 0.034
LSTM + Attention	0.604 ± 0.009	0.743 ± 0.007
HAN	0.606 ± 0.007	0.715 ± 0.013
RCNN	0.620 ± 0.003	0.744 ± 0.010
BERT	0.616 ± 0.022	0.734 ± 0.047
XLNet	0.611 ± 0.007	0.664 ± 0.017
ClinicalBERT	0.648 ± 0.011	0.774 ± 0.006
Clinical XLNet	$\mathbf{0.663 \pm 0.011}$	$\mathbf{0.779 \pm 0.006}$
Clinical XLNet- mean	0.656 ± 0.003	0.773 ± 0.003

cess was conducted on a server with 2 Intel Xeon E5-2670v2 2.5GHZ CPUs, 128GB RAM, and 2 NVIDIA Tesla P40 GPUs.

Baselines. We conducted a thorough set of experiments with several popular baselines:

- **LSTM** (Hochreiter and Schmidhuber, 1997) is the classic language modeling method that uses long term document memory.

- **LSTM + Attention** adds an attention layer on top of the sequence output of LSTM hidden layers.

- **Hierarchical Attention Networks (HAN)** (Hochreiter and Schmidhuber, 1997) is a hierarchical LSTM designed specifically for document level text classification.

---

[1]Code and models are available at https://github.com/lindvalllab/clinicalXLNet.

- **Recurrent Convolutional Neural Network (RCNN)** (Yang et al., 2016) uses a recurrent structure on the classic CNN network to capture contextual information as far as possible.

- **BERT** (Devlin et al., 2019) uses transformer encoder with the same pre-train and finetune procedure as Clinical XLNet.

- **XLNet** (Yang et al., 2019) is Clinical XLNet without pre-training on clinical text.

- **ClinicalBERT** (Huang et al., 2019; Alsentzer et al., 2019) further pre-trains on BERT using MIMIC-III notes dataset.

- **Clinical XLNet-mean** is an ablation study that uses the average of the prediction scores from each note, instead of the bidirectional LSTM layers.

Note that for BERT, XLNet, and ClinicalBERT, we all attach a bidirectional-LSTM layer on top of them to leverage the sequential dimension of notes. And for ClinicalBERT, we pre-train using the same corpus as the Clinical XLNet. These steps ensure a fair comparison between note representation power. Note that it is computationally infeasible to test the ablation of pre-finetuning.

Results. Table. 2 reports the result for our prolonged mechanical ventilation and 90-days mortality tasks. Clinical XLNet achieves the best results with AUROC score of 0.663 ($\pm$ 0.011) and 0.779 ($\pm$ 0.006) for PMV and 90-days mortality respectively. From the difference between clinically pre-trained embedding Clinical XLNet & Clinical BERT and no pre-trained model BERT & XLNet, we demonstrate the necessity of pre-training on domain-specific corpus. From the difference between Clinical BERT and Clinical XLNet, we show our Clinical XLNet has better note representation. From the difference between Clinical XLNet and Clinical XLNet-mean, we see the usage of sequential modeling of the temporal dimension of notes.

5 Discussion

In this work, we propose a method for predicting prognosis based on only contextual information available from clinician notes. The proposed method is based on the recent advancements in the field of NLP. Therefore we compare it with other recently proposed methods in NLP for a fair comparison. We perform one ablation study to show the relevance of the proposed sequential modeling of the notes embedding in the time domain and we compare against several state of the art baselines which have been used extensively in the natural language domain as well as in the clinical context.

Clinical Relevance. Our work provides timely aid in clinical decision making. For a patient under the ICU observation, the clinical team could start the evaluation to consider a tracheotomy procedure as soon as 48 hours after mechanical ventilation. The time period of 48 hours was chosen in consultation with a team of clinicians. Furthermore, a predictive analytics on the prolonged mechanical ventilation for seven days or more is important for clinicians in deciding the tracheotomy decision. Besides, the doctors could reduce the risk of a burdensome procedure and treatment by considering the patient's 90-days mortality prediction. This approach assist patients and their families by providing more time to process and make a major decision.

Model Efficiency. The proposed method uses XLNet (Yang et al., 2019) which uses TransformerXL (Dai et al., 2019) as the base architecture to extract embedding from the notes. Since every set of notes require their individual embeddings, we run the base architecture for multiple runs where the number of runs is proportional to the number of notes. Therefore, obtaining the embedding for the whole sequence of notes is computationally expensive both during the training as well as inference. However, there is a recent line of work (Lan et al., 2019) which can allow executing the transformer based models at a much lesser computational cost.

Limitations and Future Work. Our proposed method only mines task relevant information from clinical notes from nurses and respiratory therapists. However, one can utilize other sources of data as well such as structured notes. While structured data are commonly used in prognostic models, our preliminary study showed that they did not improve the performance by any significant factor. One future direction could be to explore a novel architecture design that could utilize both sources of information to improve the performance further. Another future direction would be to explore ways of combining multiple sources of clinical notes such as physician notes, admission notes, and dsischarge notes.

Acknowledgement

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *NAACL Clinical NLP Workshop*.

Connors A Herridge M Marsh B Melot C et al Boles JM, Bion J. 2007. Weaning from mechanical ventilation. *Eur Respir J.*, 29:1033–10563.

Christopher E. Cox, Shannon S. Carson, George M. Holmes, Ann Howard, and Timothy S. Carey. 2004. Increase in tracheostomy for prolonged mechanical ventilation in north carolina, 1993–2002. *Critical Care Medicine*, 32(11):2219–2226.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

J. R. Le Gall. 1993. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA: The Journal of the American Medical Association*, 270(24):2957–2963.

Jean-Roger Le Gall, P. Loirat, and A. Alpcrovitch. 1986. Apache ii-a severity of disease classification system. *Critical Care Medicine*, 14(8):754.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks : the official journal of the International Neural Network Society*, 18 5-6:602–10.

Sameer Hirji, Siobhan McGurk, Spencer Kiehm, Julius Ejiofor, Fernando Ramirez-Del Val, Ahmed A Kolkailah, Natalia Berry, Piotr Sobieszczyk, Marc Pelletier, Pinak Shah, et al. 2020. Utility of 90-day mortality vs 30-day mortality as a quality metric for transcatheter and surgical aortic valve replacement outcomes. *JAMA cardiology*, 5(2):156–165.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Abhiram Mallick and Andrew R Bodenham. 2010. Tracheostomy in critically ill patients. *European Journal of Anaesthesiology*, page 1.

William T. Mcgee. 2010. Expectations and outcomes of prolonged mechanical ventilation. *Critical Care Medicine*, 38(5):1393–1394.

Yoshihiro Mise, Jean-Nicolas Vauthey, Giuseppe Zimmitti, Nathan H Parker, Claudius Conrad, Thomas A Aloia, Jeffery E Lee, Jason B Fleming, and Matthew HG Katz. 2015. 90-day postoperative mortality is a legitimate measure of hepatopancreatobiliary surgical quality. *Annals of surgery*, 262(6):1071.

Travis B. Murdoch and Allan S. Detsky. 2013. The inevitable application of big data to health care. *Jama*, 309(13):1351.

Judith Nelson, Shannon Carson, and Thomas Bice. 2015. To trach or not to trach: Uncertainty in the care of the chronically critically ill. *Seminars in Respiratory and Critical Care Medicine*, 36(06):851–858.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159.

Amy Rapsang and Devajit C. Shyam. 2014. Scoring systems in the intensive care unit: A compendium. *Indian Journal of Critical Care Medicine*, 18(4):220–228.

Mark Unroe. 2010. One-year trajectories of care and resource utilization for recipients of prolonged mechanical ventilation. *Annals of Internal Medicine*, 153(3):167.

J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. 1996. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Marya D. Zilberberg, Rose S. Luippold, Sandra Sulsky, and Andrew F. Shorr. 2008. Prolonged acute mechanical ventilation, hospital resource utilization, and mortality in the united states. *Critical Care Medicine*, 36(3):724–730.