Enhancing Bacterial Infection Prediction in Critically Ill Patients by Integrating Clinical Text

Jinghui Liu **Anthony Nguyen**
The Australian e-Health Research Centre
CSIRO
{jinghui.liu,anthony.nguyen}@csiro.au

Abstract

Bacterial infection (BI) is an important clinical condition and is related to many diseases that are difficult to treat. Early prediction of BI can lead to better treatment and appropriate use of antimicrobial medications. In this paper, we study a variety of NLP models to predict BI for critically ill patients and compare them with a strong baseline based on clinical measurements. We find that choosing the proper text-based model to combine with measurements can lead to substantial improvements. Our results show the value of clinical text in predicting and managing BI. We also find that the NLP model developed using patients with BI can be transferred to the more general patient cohort for patient risk prediction.

1 Introduction

Data-driven AI models for healthcare have much potential to facilitate clinical care, promote healthcare efficiency, and support medical research (Topol, 2019; Rajpurkar et al., 2022). An important domain of medicine that could benefit from AI is infectious disease, where AI can help better understand infections so that we can design more effective approaches to monitor, diagnose, and treat infections (Wong et al., 2023). Among the different types of infections, bacterial infection (BI) is one of the most common and is estimated to be associated with more than 13 million deaths in 2019 alone (Collaborators, 2022).

Previous works have studied various types of AI models to predict the occurrence of BI-related diseases using data from Electronic Health Records (EHR), especially sepsis (Moor et al., 2021). Meanwhile, the prediction of BI in general is less studied, whereby structured measurements were used predominantly to develop predictive models (Yang et al., 2023; Eickelberg et al., 2023). The value of clinical text in BI prediction remains unclear.

In this study, we explore the usefulness of NLP for infection-related prediction task by focusing on BI prediction in critically ill patients. We follow an existing study (Eickelberg et al., 2020) on BI prediction that relies on a range of clinical measurements as features, and we compare it with common NLP models that rely solely on routinely collected clinical text (illustrated in Figure 1). We then use the best performing text encoder to develop multimodal fusion models for BI prediction, which obtains the state-of-the-art result. Finally, we study the applicability of NLP models for mortality prediction in different patient cohorts, showing that the model trained using patients with BI is more robust to data shift.

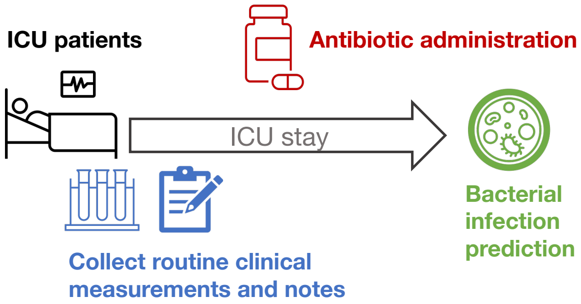
ICU patients
Antibiotic administration
ICU stay
Bacterial infection prediction
Collect routine clinical measurements and notes

Figure 1: Clinical text is integrated with clinical measurements to enhance the early prediction of bacterial infection, potentially helping inform clinical decisions regarding shortening the duration of unnecessary antibiotics to reduce risk of adverse patient outcomes and antimicrobial resistance.

2 Related Work

Many studies have developed machine learning models to predict diseases caused by bacterial infections, with urinary tract infection (Taylor et al., 2018; Dhanda et al., 2023) and sepsis (Liu et al., 2019; Moor et al., 2019) being the two most prominent examples. Early identification of these diseases is helpful, and sometimes essential, for clinicians to arrange lifesaving treatments. These studies typically use clinical measurements as features

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 118–124
November 29 - December 1, 2023 ©2023 Association for Computational Linguistics

for model development and may sometimes derive features from text as a supplement (Goh et al., 2021; Yan et al., 2022). Previous work studying BI prediction used clinical measurements (Eickelberg et al., 2020), and this was recently extended in a multicenter study (Eickelberg et al., 2023). Although text has been applied to predict specific diseases or organisms (Zhang et al., 2020), the contribution of text to BI prediction in general remains understudied.

Many previous work shows NLP models are effective for various clinical predictive tasks (Seinen et al., 2022; Liu et al., 2022a). Typical early prediction targets include patient mortality, length of stay in the hospital, readmission, diagnosis groups, or specific diseases. Multimodal fusion of different modalities in the EHR also shows promise in improving classification performance, such as combining clinical notes [1] and measurements (Deznabi et al., 2021; Soenksen et al., 2022). While previous works tend to focus on a specific type of text encoder or fusion mechanism to compare with unimodal modeling, the impact of varying these configurations on performance is not well understood.

The transferability of AI or ML models for clinical care is an important topic since many factors in healthcare can cause data shift (Finlayson et al., 2021). Applying models across different patient cohorts is also important in low-resource patient groups and to ensure fairness (Amir et al., 2021; Han et al., 2021). For example, a recent study shows that model trained in adult patients can be successfully transferred to pediatric patients (Lemmon et al., 2023). More studies are needed to understand the generalisability of models in healthcare.

3 Methods and Experiments

3.1 Task and cohort extraction

We follow Eickelberg et al. (2020) to extract a cohort of adult patients from the MIMIC-III ICU database (Johnson et al., 2016) suspected of having BI in the early phase of ICU admission. Suspicion is defined as 1) receiving at least one antibiotic within 96h after admission to the ICU and 2) having a microbiology culture tested within 24h before or after antibiotic use. For antibiotics, a duration over 96h is considered prolonged antibiotic use. For microbiology cultures, a positive culture means that a bacterial organism is detected [2]; thus, infection occurs. Unlike works focusing on specific bacteria, such as *E. coli*, we consider all possible bacteria identified from microscopy. Then, the binary classification task of BI considers prolonged antibiotic use and positive microbiologic culture as *positive* and short use and negative culture as *negative*. We follow the open source implementation to construct and process the cohort [3].

For input, we extract clinical measurements and clinical notes for patients suspected of BI. We follow Wang et al. (2020) to extract clinical measurements within the 24h data collection window from the first antibiotic dose after ICU admission. These measurements include routinely collected vital signs (such as heart rate and blood pressure) and laboratory results (such as white blood cell counts). We refer the readers to Wang et al. (2020) for a complete list of 104 clinical measurements. We did not experiment with longer windows as in (Eickelberg et al., 2020) for the purpose of this study. For clinical notes, to consider context before ICU admission, we collect all notes written before the 24th hour of ICU admission, such as those written when the patient was admitted to the hospital but not yet transferred to the ICU. We remove patients who do not have any notes recorded from the cohort. We then follow Eickelberg et al. (2020) to create train/validation/test sets with 70/10/20 ratio, where we ensure that a patient with multiple admissions appears only in one set. The statistics of the datasets are presented in Table 1.

	Train	Validation	Test
Num of cases	5937	984	2972
BI rate	19.6%	20.7%	19.6%
Mortality rate	11.7%	12.5%	10.3%
Avg num of notes	13.4	13.6	14.4
Avg num of words	4164.7	4114.2	4596.1

Table 1: Statistics of the BI cohort.

3.2 Data representation and modeling

3.2.1 Modeling clinical measurements

The structured clinical measurements are preprocessed and formatted as time series following the existing benchmark (Wang et al., 2020). We

[1] We use the terms of *clinical note* and *clinical text* interchangeably in this paper.

[2] Common contaminations are controlled by counting certain bacteria twice, i.e., *Staphylococcus*.

[3] https://github.com/geickelb/mimiciii-antibiotics-opensource

Model	AUC-ROC		AUC-PRC	
Measurement-based model	**0.772 (0.0029)**		**0.505 (0.0029)**	
Text-based models	Default ordering	Reverse ordering	Default ordering	Reverse ordering
TextCNN	0.706 (0.0041)	0.759 (0.0054)	0.346 (0.0062)	0.434 (0.0088)
BiLSTM	0.585 (0.0056)	0.646 (0.0108)	0.245 (0.0023)	0.289 (0.0093)
BERT	0.635 (0.0118)	0.717 (0.0074)	0.275 (0.0082)	0.399 (0.0145)
BERT+LSTM	0.703 (0.0099)	0.715 (0.0041)	0.337 (0.0112)	0.391 (0.0049)
Longformer	0.629 (0.0057)	0.743 (0.0026)	0.281 (0.0016)	0.437 (0.0032)

Table 2: Results of the measurement-based model and different NLP models for BI prediction. The best scores are bolded, and the second best are underscored. All scores are averaged over five runs with different random seeds.

use GRU-D in our study (Che et al., 2018), which is a strong baseline for classifying physiological time series (Rubanova et al., 2019).

3.2.2 Modeling clinical text

We consider a variety of NLP models to process clinical notes for the BI prediction task.

TextCNN: We follow the standard implementation of the classic text CNN model with multiple filters (Kim, 2014). Pretrained, in-domain word embeddings are used (Zhang et al., 2019). All notes are concatenated as a single text string as input.

BiLSTM: Previous work shows that bidirectional LSTM can be a competitive baseline even compared with more complex models for text classification (Adhikari et al., 2019). The input text is processed as for TextCNN.

BERT: We fine-tune BERT (Devlin et al., 2019) for BI classification. As pretrained BERT has an input cap of 512 tokens, the notes are concatenated and then truncated to fit this size. We use the in-domain ClinicalBERT (Alsentzer et al., 2019).

BERT+LSTM: BERT is used to encode each clinical note (first 512 tokens) and form a time series for modeling with another encoder (Zhang et al., 2020; Liu et al., 2023). We adopt this hierarchical strategy by encoding notes with ClinicalBERT to get [CLS] token representations to then model with an LSTM.

Longformer: To expand the capacity of pretrained language models, we fine-tune Longformer (Beltagy et al., 2020) with an input size of 2048 tokens. We also initialize it with in-domain pretrained weights (Li et al., 2023).

We tested two methods of ordering clinical notes. The first is the default ordering following temporal order. The other is to reverse the temporal ordering so that the most recent note appears first. Having the most updated notes appear first can be important for models with limited context length.

3.2.3 Multimodal fusion

Clinical measurements and text are combined to see if BI prediction performances can be improved. The measurements are again encoded by GRU-D. We follow previous work (Liu et al., 2023) to adopt BERT+LSTM as text encoder and then fuse with GRU-D using late fusion (Huang et al., 2020) or the attention-based fusion mechanism (Liu et al., 2023). Finally, to obtain the best result and explore whether text encoder selection matters, we select the best NLP model from the models we examined and combine it with measurement using late fusion.

3.3 Experiments

We use the area under the receiver operating curve and the precision recall curve (AUC-ROC and AUC-PRC) as metrics to evaluate the performance. We perform early stopping based on AUC-ROC (main metric) in the validation set if the score plateaus for more than five epochs for CNN and LSTM models. We tune hyperparameters for all models with grid search (see search space in the Appendix A). After finding the best configuration, the model is trained using five random seeds, whose results in the test set are averaged and presented as mean and standard deviation.

4 Results and Discussion

4.1 Modeling clinical measurement is overall better than text for BI prediction

We present the modeling results using a single input modality in Table 2. The first observation is that our implementation of GRU-D using measurements from the 24h data collection window achieves a similar performance in Eickelberg et al. (2020), where their AUC-ROC results with different classifiers range from 0.763 to 0.776, indicating that our experimental setup is consistent with previous work. We then find that the measurement-based model performs better than all the NLP models examined. This trend is similar to other clinical prediction tasks, such as mortality prediction, where structured data can outperform text (Hsu et al., 2020). This is likely because measurements can capture detailed and quantitative fast-changing physiology in patients, not consistently found in clinical notes. (Gong and Guttag, 2018).

4.2 Choice of NLP models is important for BI prediction

Nevertheless, we find text-based models can achieve competitive performances for BI prediction, especially when we reverse the order of notes. TextCNN and Longformer obtain the second best results with reversed note ordering for AUC-ROC and AUC-PRC, respectively, and approach the best results from the measurement-based model. Reverse ordering (i.e., using the lastest portions of clinical notes) brings significant benefits for models with limited context length (i.e., BERT and Longformer), which means having more sophisticated methods to select specific portions of clinical notes (Zheng et al., 2023) or remove text redundancies (Liu et al., 2022b) can potentially bring further performance boosts for BI prediction – an avenue for furture investigations.

In addition, we also observe the significant disparity between different NLP models. For example, BiLSTM obtains unexpectedly poor results compared to other methods. This may indicate that RNN is not suitable for clinical text (Boag et al., 2018) as term-level triggers may be sufficient, which can be better identified by CNN. Our results indeed show that TextCNN performs well under all settings, except when compared with Longformer under AUC-PRC. The pretrained transformer models overall underperform the simpler CNN model despite having adapted to the clinical domain and prolonged input context (i.e., Longformer). We suspect that this is because the vocabularies used by ClinicalBERT and ClinicalLongformer are not domain-specific (Koto et al., 2021) and do not handle the noise in the clinical text well. In addition, we follow Li et al. (2023) to decide the hyperparameter space when fine-tuning Longformer. It is possible that Longformer can achieve better results with more computation resources and further hyperparameter tuning. In this study, we have choosen TextCNN to balance performance and efficiency for BI prediction, and used it in combination with clinical measurements for multimodal fusion.

Model	AUC-ROC	AUC-PRC
Measurement-based model	0.772 (0.0029)	0.505 (0.0029)
Fusion with note representations encoded by BERT		
Late fusion	0.774 (0.0019)	0.508 (0.0049)
Attention-based fusion	0.781 (0.0045)*	0.508 (0.0077)
Fusion with the best text-based model		
Late Fusion	**0.799 (0.0047)***	**0.541 (0.0052)***

Table 3: BI prediction results using both measurement and text. Scores with * denote statistically significant improvement compared to measurement-based model (p-value < 0.01).

4.3 Fusion with proper NLP model improves BI prediction

Table 3 presents the results of combining measurement and text for the prediction of BI. We follow previous works to use BERT+LSTM as text encoder (Liu et al., 2023), but it provided limited benefit even with more complicated attention-based fusion mechanisms. It shows that BI prediction is different from common clinical prediction tasks in utilizing information from the two modalities. Also, the text-based BERT+LSTM alone achieves suboptimal results, which is likely the factor that limits its fusion performance. We thus select the best text encoder from Table 2 (TextCNN with reverse note ordering) and combine with measurement-based model using late fusion, which obtains significantly improved performances (p-value < 0.01, T-test). This shows that finding a proper NLP encoder for multimodal fusion can bring considerable boost to the early prediction of BI.

4.4 Bi cohort is robust to training NLP models for risk prediction

Finally, we use the Bi cohort to train an NLP model to predict in-hospital mortality and compare with another model trained using a general cohort of ICU patients, who may or may not have bacterial infection. The size of the General cohort is about 4.5 times that of the Bi cohort (more details in Appendix B). Patients in each of the train, validation and test sets of the Bi cohort appear in the corresponding set of the General cohort. We again use TextCNN with reverse ordering for model training and evaluation.

Table 4 shows the results of the mortality pre-

Model	AUC-ROC	AUC-PRC
Model trained using BI cohort		
BI test set	0.814 (0.0085)	0.377 (0.0121)
GENERAL test set	0.809 (0.0044)	0.368 (0.0106)
Model trained using GENERAL cohort		
GENERAL test set	0.893 (0.0016)	0.592 (0.0031)
BI test set	0.757 (0.0134)	0.481 (0.0241)

Table 4: The mortality prediction results on two cohorts.

diction in the two cohorts. Models trained on the BI cohort and the GENERAL cohort achieve the AUC-ROC of 0.814 and 0.893 in their corresponding in-distribution test sets. The model trained on GENERAL appears to be more capable given that it has seen more samples. We then apply these models to the test sets from the different cohorts.

Now we see that the model trained on GENERAL performs significantly worse on the BI test set (0.893 to 0.757), while the BI model maintains its performance (0.814 to 0.809). This has two implications. First, it shows that a risk prediction model trained using a general population cannot be directly applied to patients with bacterial infection (AUC-ROC drops from 0.814 to 0.757) and a dedicated model needs to be trained. This relates to the effect of data bias on subpopulations that causes models to learn shortcuts and perform differently across various groups of patients (Brown et al., 2023). Second, patients with bacterial infection turn out to be a valuable resource for training a robust risk prediction model that can be applied to a broader cohort. We consider that this finding warrants future investigation of the factors that lead to the difference and ways to develop a more transferable clinical prediction model for different groups of patients.

5 Conclusion

Clinical text can help predict BI in critically ill patients and NLP models trained using BI patients can be transported to those without BI. NLP and multimodal models can develop better data-driven strategies to stratify the risk of BI in patients, which can be compared with prompt-based large language models (LLMs) in future work. Clinical co-development will be pursued to ensure that the developed models are optimised for clinical workflow, capable of refining antibiotic therapy in the absence of test results, and have the potential to enhance antimicrobial stewardship, thereby mitigating antimicrobial resistance. In the future, we would like to investigate how text can help improve BI treatment, such as antimicrobial stewardship and predict potential antimicrobial resistance.

Acknowledgements

We thank the anonymised reviewers for their comments to help improve the manuscript during the review process.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Rethinking complex neural network architectures for document classification. In *NAACL*, pages 4046–4051.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Clinical Natural Language Processing Workshop*, pages 72–78.

Silvio Amir, Jan-Willem van de Meent, and Byron Wallace. 2021. On the impact of random seeds on the fairness of clinical classifiers. In *NAACL*, pages 3808–3823.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document transformer.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What's in a note? unpacking predictive value in clinical note representations. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:26–34.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. 2023. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature communications*, 14(1):4314.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

GBD 2019 Antimicrobial Resistance Collaborators. 2022. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 400(10369):2221–2248.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Iman Deznabi, Mohit Iyyer, and Madalina Fiterau. 2021. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4026–4031, Online. Association for Computational Linguistics.

Gurpreet Dhanda, Mirna Asham, Denton Shanks, Nicole O'Malley, Joel Hake, Megha Teeka Satyan, Nicole T Yedlinsky, and Daniel J Parente. 2023. Adaptation and external validation of pathogenic urine culture prediction in primary care using machine learning. *Annals of family medicine*, 21(1):11–18.

Garrett Eickelberg, L Nelson Sanchez-Pinto, and Yuan Luo. 2020. Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults. *Journal of biomedical informatics*, 109:103540.

Garrett Eickelberg, Lazaro Nelson Sanchez-Pinto, Adrienne Sarah Kline, and Yuan Luo. 2023. Transportability of bacterial infection prediction models for critically ill patients. *Journal of the American Medical Informatics Association: JAMIA*.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. 2021. The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3):283–286.

Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. 2021. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications*, 12(1):711.

Jen J Gong and John V Guttag. 2018. Learning to summarize electronic health records using Cross-Modality correspondences. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 551–570. PMLR.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *EACL*, pages 2760–2765.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the value of information in medical notes. In *Findings of EMNLP 2020*, pages 2062–2072.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. 2020. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3:136.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, 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.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective Domain-Specific vocabulary initialization. In *EMNLP*, pages 10660–10668.

Joshua Lemmon, Lin Lawrence Guo, Ethan Steinberg, Keith E Morse, Scott Lanyon Fleming, Catherine Aftandilian, Stephen R Pfohl, Jose D Posada, Nigam Shah, Jason Fries, and Lillian Sung. 2023. Self-supervised machine learning using adult inpatient data produces effective models for pediatric clinical prediction tasks. *Journal of the American Medical Informatics Association: JAMIA*.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association: JAMIA*, 30(2):340–347.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022a. Improving text-based early prediction by distillation from privileged Time-Series text. In *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, pages 73–83, Adelaide, Australia. Australasian Language Technology Association.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022b. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2023. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of biomedical informatics*, 145:104466.

Ran Liu, Joseph L Greenstein, Stephen J Granite, James C Fackler, Melania M Bembea, Sridevi V Sarma, and Raimond L Winslow. 2019. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Scientific reports*, 9(1):6145.

Michael Moor, Max Horn, Bastian Rieck, Damian Roqueiro, and Karsten Borgwardt. 2019. Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time warping. In

Proceedings of the 4th Machine Learning for Healthcare Conference, volume 106 of *Proceedings of Machine Learning Research*, pages 2–26. PMLR.

Michael Moor, Bastian Rieck, Max Horn, Catherine R Jutzeler, and Karsten Borgwardt. 2021. Early prediction of sepsis in the ICU using machine learning: A systematic review. *Frontiers of medicine*, 8:607952.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature medicine*, 28(1):31–38.

Yulia Rubanova, Ricky T Q Chen, and David K Duvenaud. 2019. Latent ordinary differential equations for Irregularly-Sampled time series. In *NeurIPS*, volume 32.

Tom M Seinen, Egill A Fridgeirsson, Solomon Ioannou, Daniel Jeannetot, Luis H John, Jan A Kors, Aniek F Markus, Victor Pera, Alexandros Rekkas, Ross D Williams, Cynthia Yang, Erik M van Mulligen, and Peter R Rijnbeek. 2022. Use of unstructured text in prognostic clinical prediction models: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 29(7):1292–1302.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. 2022. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149.

R Andrew Taylor, Christopher L Moore, Kei-Hoi Cheung, and Cynthia Brandt. 2018. Predicting urinary tract infections in the emergency department with machine learning. *PloS one*, 13(3):e0194085.

Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56.

Shirly Wang, Matthew B A McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *ACM Conference on Health, Inference, and Learning*, pages 222–235.

Felix Wong, Cesar de la Fuente-Nunez, and James J Collins. 2023. Leveraging artificial intelligence in the fight against infectious diseases. *Science*, 381(6654):164–170.

Melissa Y Yan, Lise Tuset Gustad, and Øystein Nytrø. 2022. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 29(3):559–575.

Ying Yang, Yi-Min Wang, Chun-Hung Richard Lin, Chi-Yung Cheng, Chi-Ming Tsai, Ying-Hsien Huang, Tien-Yu Chen, and I-Min Chiu. 2023. Explainable deep learning model to predict invasive bacterial infection in febrile young infants: A retrospective study. *International journal of medical informatics*, 172:105007.

Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. 2020. Time-Aware transformer-based network for clinical notes series prediction. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of Machine Learning Research*, pages 566–588. PMLR.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, 6(1):52.

Hongyi Zheng, Yixin Zhu, Lavender Jiang, Kyunghyun Cho, and Eric Oermann. 2023. Making the most out of the limited context length: Predictive power varies with clinical note type and note section. In *The 61st ACL (Student Research Workshop)*, pages 104–108.

A Hyperparamter Tuning

For TextCNN, BiLSTM, and BERT+LSTM models, we sweep through the space: number of RNN hidden state/CNN filter number $\in [128, 256, 512]$; dropout rate $\in [0.2, 0.4, 0.6]$; weight decay $\in [0, 0.01]$; learning rate $\in [1e-3, 1e-4]$. The batch size is kept as 32. For BERT fine-tuning, we explore epoch $\in [3, 5, 10]$ and learning rate $\in [2e-5, 3e-5, 5e-5]$. For Longformer fine-tuning, we explore learning rate $\in [1e-5, 2e-5, 5e-5]$ and kept epoch as 5 to save computation. The batch size for the two models is kept as 16 using gradient accumulation.

B Constructing GENERAL Cohort

We follow the criteria in previous work (Hsu et al., 2020; Harutyunyan et al., 2019) to select this cohort of patients and use notes charted before 24 hours of admission to the ICU as input, the same as in the BI cohort. There are three criteria for selection: 1) adult patients, 2) no repeated ICU admissions, and 3) hospital discharge time is at least 30 hours away from ICU admission. Table 5 shows the statistics of the cohort.

	Train	Validation	Test
Num of cases	30162	4475	10320
Mortality rate	10.2%	10.4%	9.6%
Avg num of notes	8.7	8.6	8.7
Avg num of words	2440.3	2432.8	2468.4

Table 5: GENERAL cohort statistics.