Benchmarking Foundation Models For Antibiotic Susceptibility Prediction

Helio Halperin[*]
heliohalperin@gmail.com

Yanai Halperin *
yanaihalperin@gmail.com

Simon A. Lee [†]
simonlee711@g.ucla.edu

Jeffrey N. Chiang [† ‡]
njchiang@g.ucla.edu

Abstract

The rise of antibiotic-resistant bacteria has been identified as a critical global healthcare crisis that compromises the efficacy of essential antibiotics. This crisis is largely driven by the inappropriate and excessive use of antibiotics, which leads to increased bacterial resistance. In response, clinical decision support systems integrated with electronic health records (EHRs) have emerged as a promising solution. These systems use machine learning models to improve antibiotic stewardship by providing actionable, data-driven insights. This study therefore evaluates pre-trained foundation models for predicting antibiotic susceptibility, using several open-source models available on the Hugging Face platform. Despite the abundance of models and ongoing advancements in the field, a consensus on the most effective model for encoding clinical knowledge remains unclear.

1 Introduction

The Centers for Disease Control and Prevention (CDC) identifies antibiotic-resistant bacteria as a critical global health crisis, threatening the efficacy of essential antibiotics [Ventola, 2015, Lushniak, 2014, Nature, 2013, Piddock, 2012, Bartlett et al., 2013]. This crisis stems from the improper and excessive use of antibiotics, enhancing bacterial resistance Viswanathan [2014], Sengupta et al. [2013]. These resistant strains add significant clinical and financial burdens globally.

Clinical decision support systems, part of health information technology integrated with EHRs [Hoerbst and Ammenwerth, 2010], aid healthcare professionals in decision-making [Wu et al., 2010, Sutton et al., 2020] and enhance antibiotic stewardship through data-driven insights Lee et al. [2024a]. In this study, we assess pre-trained language models, or foundation models, to predict antibiotic susceptibility in STAPH infection cases. Leveraging research that integrates tabular EHR data with language models Lee et al. [2024b], we aim to identify the most effective model, addressing a critical healthcare challenge amidst the ongoing release and development of new foundation models.

2 Methodology

In this study, we are interested in utilizing Electronic Health Records in text form to predict the susceptibility of several antibiotics using recent pre-trained foundation models. Identifying the best model that encodes this clinical information is crucial in healthcare, as improved performance metrics lead to better decision support systems that directly benefit patients.

[*]Santa Monica High School
[†]Department of Computational Medicine, UCLA
[‡]Department of Neurosurgery, UCLA

Preprint. Under review.

Table 1: Foundation Models used in our Benchmarking Study

Name/HuggingFace Model Card	Source	Name/HuggingFace Model Card	Source
pritamdeka/BioBert-PubMed200kRCT	[Deka et al., 2022]	distil-bert	[Sanh et al., 2019]
emilyalsentzer/Bio_ClinicalBERT	[Alsentzer et al., 2019]	UFNLP/gatortron-base	[Yang et al., 2022]
EMBO/bio-lm	———	michiyasunaga/LinkBERT-large	[Yasunaga et al., 2022]
allenai/biomed_roberta_base	[Gururangan et al., 2020]	Charangan/MedBERT	[Vasantharajan et al., 2022]
EMBO/BioMegatron345mUncased	[Shin et al., 2020]	NeuML/pubmedbert-base	———
bionlp/bluebert_pubmed_mimic_uncased	[Peng et al., 2019]	StanfordAIMI/RadBERT	[Chambon et al., 2022]
medicalai/ClinicalBERT	[Wang et al., 2023]	allenai/scibert_scivocab_uncased	[Beltagy et al., 2019]

2.1 Data Source: Electronic Health Records

Electronic health records (EHRs) are digital versions of patients' medical histories, accessible across various healthcare settings. We use the publicly available MIMIC-IV database from the Medical Information Mart for Intensive Care-IV, a popular resource for AI healthcare research Johnson et al. [2023]. This database allows us to gather and select predictors and targets for our study.

Our goal is to predict susceptibility to eight antibiotics—Clindamycin, Erythromycin, Gentamicin, Levofloxacin, Oxacillin, Tetracycline, Trimethoprim, and Vancomycin—as per protocols in [Lee et al., 2024a]. We focus on STAPH infection patients from the Emergency Department (ED) to form our study group, identifying 4,161 patients with 5,976 prescriptions, all with clear data labels. We use six EHR tables tied to ED care, covering arrival, medication history, vitals, in-stay medications (Pyxis), diagnoses, and triage.

2.2 Feature Engineering: Generating Text from EHR

EHR in its canonical tabular form aren't suitable for language models, which are designed to process natural language. These models struggle with structured data like CSVs and JSONs, typical in EHRs, without extra programming to understand these formats.

To bridge this gap, we adopt an EHR feature engineering method from [Lee et al., 2024b] that converts tables into text using text template strategies Hegselmann et al. [2023], Ono and Lee [2024]. This approach, as Lee et al. found, preserves the clinical data's integrity better than traditional methods that rely solely on numerical transformations.

To describe the conversion of tabular data to text, we model each table row R as a set of attributes $a_1, a_2, \ldots, a_n$. A function f maps these attributes to a text template T, forming the text representation T_R:

$$T_R = f(R) = f(a_1, a_2, \ldots, a_n)$$

Function f concatenates attribute values into a structured text, preserving the semantic integrity of the data for language model processing. This method is applied to each table in the EHR and then combined to make a patient paragraph.

2.3 Experimental Task

In Section 2.1, we defined the antibiotics of interest for our study. We treat each antibiotic individually as a binary classification problem, where the model predicts whether a patient is susceptible or not to that particular antibiotic. We will measure the Area Under the Receiver Operating Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) for each foundation model in our benchmark.

2.4 Pre-trained Language Models

In our benchmarking, we evaluate foundation models from Table 1, specifically choosing open-source, small-sized (around 100M parameters) models from Hugging Face [Wolf et al., 2019] due to hardware constraints and practicality in healthcare settings. Predominantly, these are small BERT-derived models for sequence classification. We excluded other EHR models previously assessed in [Lee et al., 2024a], where BERT-based models outperformed EHR-specific and table based models.

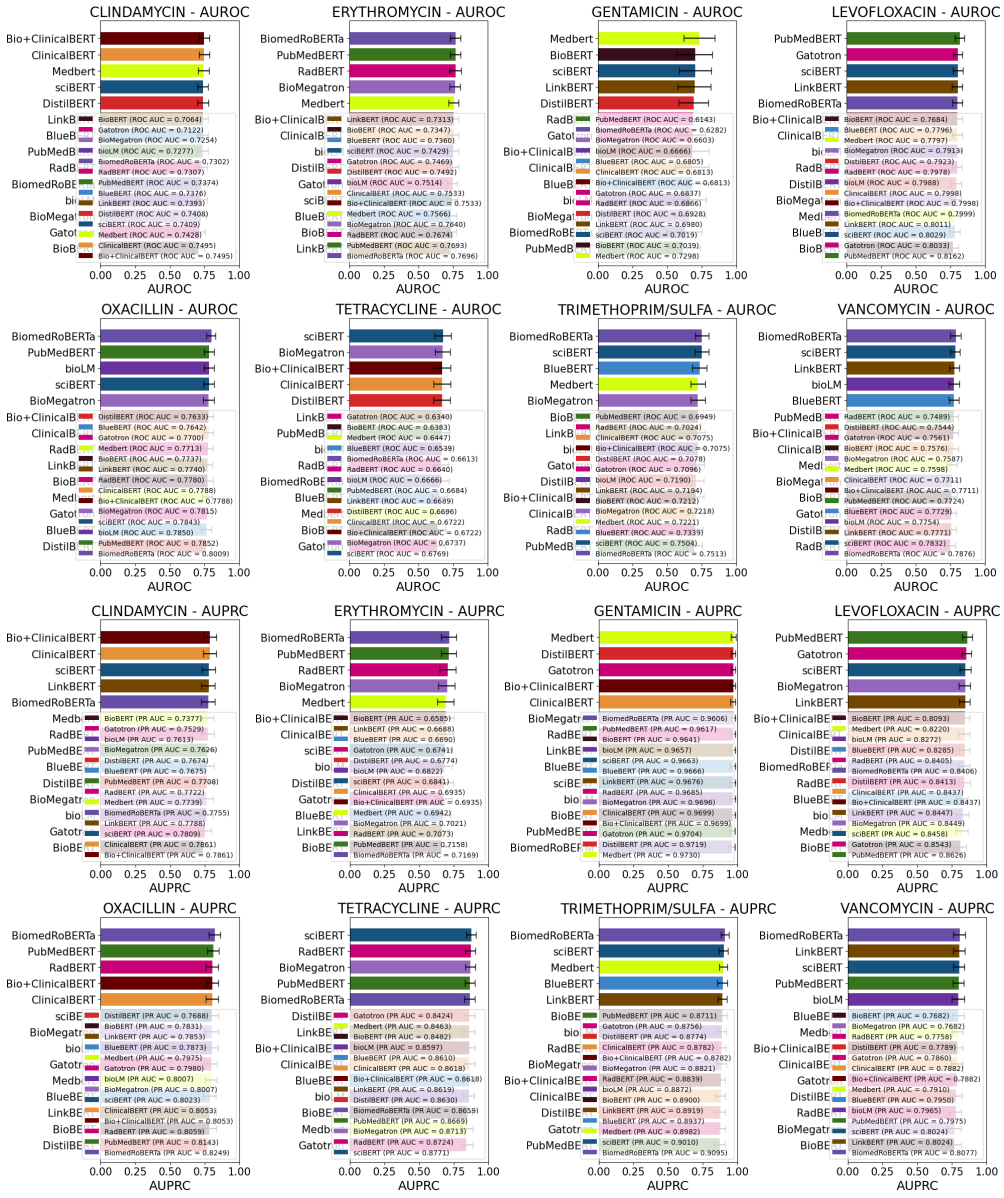

Figure 1: Performance comparison of Foundation models in predicting antibiotic susceptibility for AUROC (top) and AUPRC (bottom) in descending order.

2.5 Modeling Setup

In this work, we generate embeddings from our foundation models, which are vector representations of our text. We freeze the parameters to this model meaning the foundation model's parameters do not change. From here, we train a light gradient boosting machine (LGBM) as our classifier previously done by [Chen et al., 2024]. We train 8 independent classifiers for each antibitoic in our study to obtain our results.

3 Results

In Figures 1 we present the AUROC and AUPRC graphs for each corresponding antibiotic. We order the bars from best to worst to indicate the winning model for each antibitoic. However, to understand

Table 2: Average Rank of Models in AUROC and AUPRC

Model Name	AUROC Rank	AUPRC Rank
sciBERT	**3.75**	**4.375**
BiomedRoBERTa	5.25	4.75
ClinicalBERT	6.0	5.75
LinkBERT	6.625	7.25
PubMedBERT	6.75	5.875
Bio+ClinicalBERT	7.0	6.75
Medbert	7.125	7.25
BioMegatron	7.625	7.25
bioLM	7.625	9.25
DistilBERT	8.875	9.125
RadBERT	8.875	6.375
BlueBERT	9.25	9.25
Gatotron	9.875	9.375
BioBERT	10.375	12.375

which model was the best performing among these two metrics on average, we also calculate the average ranks of each methodology per metric in Table 2.

4 Discussion

4.1 BiomedRoBERTa has the most top metrics, but sciBERT performs best on average

From Figure 1, our results are nuanced, with different models performing best depending on the antibiotic. Bio_ClinicalBERT, BiomedRoBERTa, MedBERT, PubMedBERT, and sciBERT are the five models that excel in at least one antibiotic. Based on the graph alone, we observe that BiomedRoBERTa performs best for four out of eight of the antibiotics. However, when examining the average ranks in Table 2, it appears that sciBERT may be the most consistently high-performing model in our benchmarking study indicated by its average rank at 3.75 (AUROC) and 4.375 (AUPRC).

4.2 Foundation Models can help with Antibiotics Recommendations

From this study, we can see that foundation models are capable data-driven techniques to help and prescribe the correct antibiotics to patients. These models leverage extensive data about a patient to predict antibiotic susceptibility, providing vital support in clinical decision-making processes.

5 Conclusion

In this work, we conducted a benchmarking study to evaluate various foundation models and their ability to predict antibiotic susceptibility. This effort supports the broader cause of using data-driven solutions to combat the global health crisis of antibiotic resistance. Our findings indicate that BiomedRoBERTa performed best for four of the eight antibiotics studied. However, when analyzing the average ranks, sciBERT emerged as the consistently superior model.

A significant insight from our study is that clinical decision support may require extensive benchmarking for specific prediction tasks. We believe this is essential because no single model excelled across all metrics, despite being labeled as “foundation models." This observation underscores the need for future studies to assess the reliability of research papers claiming to develop the "best" foundation model or whether the model was optimized to excel at a specific task.

Limitations Our work only discusses encoder based foundation models and future studies need to be done to assess recent decoder based foundation models.

Data & Code MIMIC IV ED data is available at https://physionet.org/content/mimic-iv-ed/2.2/. Our code is available at https://github.com/antibiotics-fm-benchmark.

Acknowledgements We thank our mentor Simon Lee, the professor Jeffrey N. Chiang for their mentorship. We additionally thank the big summer program at UCLA for allowing us to participate in research.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

John G Bartlett, David N Gilbert, and Brad Spellberg. Seven ways to preserve the miracle of antibiotics. *Clinical infectious diseases*, 56(10):1445–1450, 2013.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Pierre Chambon, Tessa S. Cook, and Curtis P. Langlotz. Improved fine-tuning of in-domain transformer model for inferring covid-19 presence in multi-institutional radiology reports. *Journal of Digital Imaging*, 2022. doi: 10.1007/s10278-022-00714-8.

Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Reisler, David E Kim, and Pranav Rajpurkar. Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine. *Advances in Neural Information Processing Systems*, 36, 2024.

Pritam Deka, Anna Jurek-Loughrey, et al. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer, 2022.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, 2020.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.

Alexander Hoerbst and Elske Ammenwerth. Electronic health records. *Methods of information in medicine*, 49(04):320–336, 2010.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Simon A. Lee, Trevor Brokowski, and Jeffrey N. Chiang. Enhancing antibiotic stewardship using a natural language approach for better feature representation, 2024a.

Simon A Lee, Sujay Jain, Alex Chen, Arabdha Biswas, Jennifer Fang, Akos Rudas, and Jeffrey N Chiang. Multimodal clinical pseudo-notes for emergency department prediction tasks using multiple embedding model for ehr (meme). *arXiv preprint arXiv:2402.00160*, 2024b.

Boris D Lushniak. Antibiotic resistance: a public health crisis. *Public Health Reports*, 129(4): 314–316, 2014.

E Nature. The antibiotic alarm. *Nature*, 495(7440):141, 2013.

Kyoka Ono and Simon A Lee. Text serialization and their relationship with the conventional paradigms of tabular machine learning. *arXiv preprint arXiv:2406.13846*, 2024.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65, 2019.

Laura JV Piddock. The crisis of no new antibiotics—what is the way forward? *The Lancet infectious diseases*, 12(3):249–253, 2012.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Saswati Sengupta, Madhab K Chattopadhyay, and Hans-Peter Grossart. The multifaceted roles of antibiotics and antibiotic resistance in nature. *Frontiers in microbiology*, 4:47, 2013.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. Biomegatron: larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, 2020.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

Charangan Vasantharajan, Kyaw Zin Tun, Ho Thi-Nga, Sparsh Jain, Tong Rong, and Chng Eng Siong. Medbert: A pre-trained language model for biomedical named entity recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1482–1488, 2022. doi: 10.23919/APSIPAASC55919.2022.9980157.

C Lee Ventola. The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and therapeutics*, 40(4):277, 2015.

VK Viswanathan. Off-label abuse of antibiotics by bacteria. *Gut microbes*, 5(1):3–4, 2014.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48(6):S106–S113, 2010.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*, 2022.