Using Foundation Models to Prescribe Patients Proper Antibiotics

Simon A. Lee [1], **Helio Halperin** [1,2], **Yanai Halperin** [1,2], **Trevor Brokowski**[3,4], **Jeffrey N. Chiang**[1,5]

[1]Department of Computational Medicine, UCLA
[2]Santa Monica High School
[3]Yale School of Medicine
[4]Biomedical Informatics & Data Science, Yale University
[5]Department of Neurosurgery, UCLA

Abstract

The rise of antibiotic-resistant bacteria presents a significant global health threat by reducing the effectiveness of essential treatments. This study evaluates the potential of clinical decision support systems powered by biomedical language foundation models to enhance antibiotic stewardship using electronic health records (EHRs). We test several state-of-the-art models, focusing on predicting whether each of eight different antibiotics will be effective for an individual patient. Additionally, we emphasize interpretability, aiming to understand how the models make decisions, where they excel, and where they fall short. Unlike previous research, which primarily benchmarks accuracy metrics, we provide insights into both the successes and limitations of these models, offering clinical and non-clinical experts a clearer understanding of their current state and reliability. These findings highlight the potential of AI systems to combat this global health threat, as well as the need for further improvements to address the limitations of existing models. We hope this work offers valuable guidance for improving AI-driven decision support systems and leveraging these advanced models for other clinical applications.

Code — https://github.com/Simonlee711/antibiotics-fm-benchmark

Datasets — https://physionet.org/content/mimiciv/3.1/

Introduction

The Centers for Disease Control and Prevention (CDC) identifies antibiotic-resistant bacteria as a critical global health challenge, undermining the effectiveness of traditional antibiotics (Ventola 2015; Golkar, Bagasra, and Pace 2014; Gould and Bal 2013; Sengupta, Chattopadhyay, and Grossart 2013; Nature 2013; Lushniak 2014). The primary drivers of this resistance include the misuse and overuse of antibiotics, which promotes resistance through repeated exposure (Viswanathan 2014; Read and Woods 2014). Additionally, the decline in new antibiotic development, attributed to high costs and regulatory hurdles, exacerbates this issue (Piddock 2012). Consequently, once-manageable infections are now more difficult to treat, increasing hospital durations and healthcare costs (Bartlett, Gilbert, and Spellberg 2013).

The proliferation of multidrug-resistant pathogens, termed "superbugs", further aggravates the crisis, rendering many conventional treatments ineffective (Adegoke et al. 2016; Alpert 2017). Increasingly common multidrug-resistant strains, such as Staphylococcus aureus (MRSA), Escherichia coli, and Klebsiella pneumoniae, lead to severe infections with higher mortality rates (Warnke et al. 2013; De Kraker et al. 2011; Gandra et al. 2019). These pathogens significantly challenge infection management during medical procedures like surgeries and cancer treatments, which depend on effective antibiotics (Jones, Bunn, and Bell-Syer 2014; Gao et al. 2020). The rising prevalence of superbugs highlights the critical need for global strategies, enhanced surveillance, and innovative treatments to mitigate resistance spread.

To address this challenge, this study introduces and evaluates a clinical decision support framework utilizing biomedical language foundation models to improve antibiotic stewardship. By using pretrained language models, the framework analyzes extensive patient data to provide real-time adherence to antibiotic guidelines (Hoerbst and Ammenwerth 2010; Kohli and Tan 2016).

Unlike previous research that primarily showcases accuracy metrics, this paper conducts a comprehensive evaluation of various models, focusing on their predictive performance and interpretability. The evaluation assesses the models' effectiveness across eight different antibiotics, quantifying the accuracy of prescriptions and identifying the rates of incorrect and missed prescriptions. This analysis highlights the practical reliability of these models in clinical settings. The findings provide critical insights for clinicians and researchers, delineating the current capabilities and limitations of these models. By elucidating these models' strengths and weaknesses, this research contributes to the global initiative to combat antibiotic resistance through an foundation model based framework.

Contributions

- This study introduces the use of advanced biomedical language models to enhance antibiotic prescribing practices and adherence to stewardship guidelines.
- It evaluates the predictive accuracy and interpretability of these models, specifically assessing their performance

in prescribing eight different antibiotics and identifying areas of strength and limitation.

- The work shares the number and types of errors giving insights into the models reliability.

Related Works

Clinical Outcomes Prediction using AI

The use of Artificial Intelligence (AI) to predict clinical outcomes is a well-established practice (Magrabi et al. 2019; Shortliffe and Sepúlveda 2018). However, these systems entail more than the mere application of computational algorithms to existing datasets (Xie et al. 2022). Initially, the focus was on traditional tabular models such as logistic regression (Nick and Campbell 2007), random forests (Breiman 2001), and gradient boosting methods (Chen and Guestrin 2016; Ke et al. 2017; Prokhorenkova et al. 2018). These methods, known for their relative simplicity and interpretability, particularly in tree-based models, establish decision boundaries that guide predictions through deterministic categories and are valued in clinical studies for their transparency and the clarity they provide in understanding model decisions.

The introduction of deep learning has significantly transformed this field, as researchers now use large-scale models to predict a range of clinical outcomes (Miotto et al. 2018; Choi et al. 2017; Li, Huang, and Zitnik 2022). Despite their complexity and the "black-box" nature of their decision-making processes, these models are favored for their ability to handle high-dimensional data and approximate complex functions with remarkable sophistication (Nielsen 2016; Lu et al. 2021; Lu and Lu 2020). They are applied in various domains, including medical image interpretation, drug discovery and delivery, diagnosis, and prognosis (Tajbakhsh et al. 2020; Mullowney et al. 2023; Farnoud et al. 2022; Ávila-Jiménez et al. 2024; Kumar et al. 2024; Khalighi et al. 2024), among others (Al Kuwaiti et al. 2023; Lee, Brokowski, and Chiang 2024; Idowu et al. 2023).

Recent studies have examined whether language models trained on general or biomedical scientific text can accurately predict clinical outcomes directly from patient records. Examples include works such as (Gupta et al. 2022), Gatortron (Yang et al. 2022), MIMIC-IV-Ext (Hager, Jungmann, and Rueckert), (Yang et al. 2023a), MEME (Lee et al. 2024), CliBench (Ma et al. 2024), (Lee and Lindsey 2024) (Hager et al. 2024), and (Li et al. 2024). These methods rely on representing EHR as text, use discharge summaries, or perform feature engineering to make tabular EHR data fields compatible with emerging language model technologies.

Measuring Susceptibility to Antibiotics Using AST

Biomedical Language Models

Biomedical language models like BioBERT (Lee et al. 2020) and ClinicalBERT (Alsentzer et al. 2019) are specialized adaptations of BERT (Devlin 2018), designed for healthcare applications. BERT, a transformer-based (Vaswani 2017) model, captures the context of words bidirectionally, enhancing its effectiveness for complex language patterns. Originally pretrained on general text such as Wikipedia and BookCorpus, BERT is fine-tuned for specific tasks like named entity recognition (NER), text classification, and question answering. However, the unique vocabulary and syntax of medical language necessitate domain-specific pretraining for clinical applications.

Biomedical language models extend BERT's capabilities by further pretraining on biomedical texts, including PubMed abstracts (Gu et al. 2020), full-text articles (Beltagy, Lo, and Cohan 2019), and clinical notes (Alsentzer et al. 2019). These models are adept at capturing intricate relationships in medical language, making them suitable for clinical concept extraction and patient outcome prediction. The primary data sources for these models are electronic health records (EHRs) and clinical text data, which include both structured information (e.g., lab values, medication records) and unstructured text such as discharge summaries and radiology reports. Together, these data sources provide a comprehensive set of medical terminologies and contextual information essential for learning medical language representations.

To utilize EHR data, structured information is transformed into tokenized text that models can process. For example, an EHR dataset containing patient records with n features (e.g., demographics, diagnoses, lab results) represents each patient i with a feature vector $x_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$. Each feature x_{ij} is converted into a text token, creating a sequence $T_i = [t_{i1}, t_{i2}, \ldots, t_{in}]$, which is then tokenized using methods like WordPiece (Song et al. 2020), producing a sequence of subword tokens $S_i = [s_{i1}, s_{i2}, \ldots, s_{im}]$.

These tokens are embedded into a high-dimensional space using BERT's embedding layer, mapping each token s_{ij} to a d-dimensional vector $e_{ij} \in \mathbb{R}^d$, where d is typically 768 for BERT-base. The sequence of embedding vectors $E_i = [e_{i1}, e_{i2}, \ldots, e_{im}]$ is processed through multiple transformer layers to capture contextual information, resulting in final contextualized embeddings $C_i = [c_{i1}, c_{i2}, \ldots, c_{im}]$. Each contextualized embedding $c_{ij} \in \mathbb{R}^d$ encapsulates the semantic meaning of the token s_{ij} within the sequence, enabling these embeddings to be used for downstream tasks such as similarity search or patient outcome prediction.

Building on BERT's success in natural language processing, numerous studies have explored its application in the biomedical field. Early models like BEHRT (Li et al. 2020) and MedBERT (Rasmy et al. 2021) demonstrated the potential of adapting BERT's architecture to large-scale patient data by encoding sequences of diagnoses, prescriptions, and laboratory values as tokens. Recent advancements have expanded the range of features integrated into these models, with examples like ExBEHRT (Rupp, Peter, and Pattipaka 2023), IRENE (Zhou et al. 2023), and M-BioBERTa (Antal et al. 2024) enhancing contextual embedding processes. Additionally, models such as TransformEHR (Yang et al. 2023b), Gatortron (Yang et al. 2022), and CLMBR (Wornow et al. 2023) demonstrate that transformer decoders focusing on forward-directional attention can further improve predictive performance on tasks aimed at forecasting future medical events.

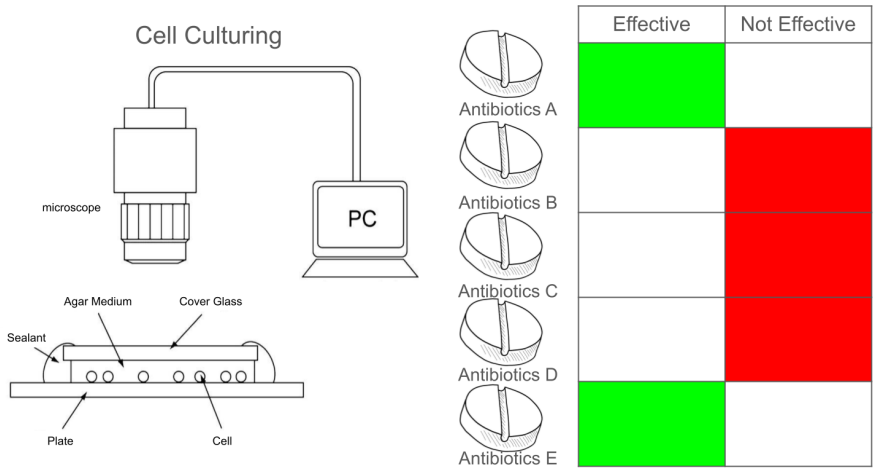

Figure 1: Antimicrobial Susceptibility Testing (AST). This figure demonstrates the experimental setup of how susceptibility to antibiotics is determined as "Susceptible", "Intermediate", or "Resistant".

Data

This section explains the data source and modeling framework used in this study, which are crucial for defining and contextualizing the problem addressed. It details the electronic health record (EHR) data used, including the origin of the data, the criteria for cohort selection, and how these elements contribute to and support the overall modeling approach.

Data Source

The MIMIC-IV-ED dataset [1] (Johnson et al. 2023) is a comprehensive, publicly available resource that contains detailed records of emergency department (ED) admissions at Beth Israel Deaconess Medical Center, spanning the years 2011 to 2019. This dataset encompasses a total of 448,972 ED stays, providing rich information on vital signs, triage assessments, medication administration, and discharge diagnoses. Such extensive and granular data make the MIMIC-IV-ED dataset a valuable resource for developing and evaluating models in clinical research, particularly in the context of patient outcomes and treatment optimization.

Cohort Selection

For this study, we focused on emergency department (ED) patients with presumed Staphylococcus aureus (STAPH) infections (Tenover and Gorwitz 2006). Inclusion criteria included positive cultures and relevant diagnostic codes. From this, we defined a cohort by filtering for cases with available antimicrobial susceptibility testing (AST) results, focusing on Staphylococcus aureus to precisely model the problem and use susceptibility labels as the target outcome.

Our objective was to predict susceptibility to eight antibiotics—Clindamycin, Erythromycin, Gentamicin, Levofloxacin, Oxacillin, Tetracycline, Trimethoprim, and Vancomycin—according to the protocols in (Lee, Brokowski, and Chiang 2024). We identified 4,161 patients with Staphylococcus aureus infections in the ED, which corresponded to 5,976 antibiotic prescriptions with clearly labeled data. Further details about the cohort and antibiotic prevalence are provided in Table 1.

Measuring a patient's susceptibility to antibiotics involves antimicrobial susceptibility testing (AST) (Reller et al. 2009; Hindler and Munro 2010; Lalitha et al. 2004). This laboratory procedure requires obtaining a bacterial sample from the patient—from sites such as blood, urine, or tissue—and culturing it in a controlled environment to promote bacterial growth. The bacteria are then exposed to various antibiotics to assess which are effective at inhibiting growth or eradicating the bacteria (Figure 1). This testing is typically conducted using standardized methods like disk diffusion (Jorgensen and Turnidge 2015) or broth microdilution (Thornsberry and Mcdougal 1983), which evaluate bacterial growth in the presence of antibiotics at various concentrations. Results are categorized as "Susceptible," "Intermediate," or "Resistant," according to thresholds set by clinical guidelines. For our modeling, we designated "susceptible" as indicating an effective antibiotic, while "intermediate" and "resistant" were considered indicators of ineffective antibiotics.

AST results are crucial for clinicians in selecting the most appropriate antibiotic treatment. However, a significant limitation of this method is the time required to culture cells, which can delay treatment for patients requiring immediate care.

Opportunities for AI Integration

In addition to using antimicrobial susceptibility testing (AST) results, the extensive structured data in electronic health records (EHRs) provides an opportunity for AI models to predict patient susceptibility before AST results are available. By analyzing a patient's medical history, AI could facilitate early intervention with effective antibiotic treatments, thereby accelerating clinical decision-making and

[1] https://physionet.org/content/mimic-iv-ed/2.2/

Table 1: MIMIC IV Cohort Data Overview and Antibiotic Prevalence

Cohort Data Overview					**Antibiotic Prevalence**			
Description	Category	Train	Test	Totals	Antibiotic	Train	Test	Total Prevalence (%)
Prescription, n	Total	4803	1173	5976	Clindamycin	2645	624	**54.69%**
Unique ID, n	Total	3283	878	4161	Erythromycin	2626	639	**54.59%**
Age Mean (SD)		59 (17)	58 (17)		Gentamicin	4549	1127	**94.89%**
Sex	Male	1942	527	2469	Levofloxacin	2866	715	**60.00%**
Race/Ethnicity %	White	2212	583	2795	Oxacillin	2702	667	**56.32%**
	Black	416	119	535	Tetracycline	3747	909	**76.57%**
	Other	567	156	497	Trimethoprim/sul	3671	908	**71.66%**
	Asian	88	20	108	Vancomycin	2529	611	**52.53%**

potentially reducing treatment delays. Utilizing a patient's medical history to predict suitable antibiotics at the time of hospital admission through AI could streamline the treatment process and enhance patient treatment optimizations.

Methods

Text-based Electronic Health Records

Electronic Health Records (EHR) in the MIMIC database are organized into multiple tables, which store varied aspects of a patient's medical history such as demographics, diagnostic history, and medications. This fragmentation presents significant challenges in data modeling and integration for machine learning applications. Traditional models tailored for tabular data, such as boosting methods including CatBoost and XGBoost, often struggle with the effective handling of categorical and free-text fields. Moreover, feature engineering techniques like one-hot encoding can greatly increase dimensionality, leading to sparsity and multicollinearity issues.

To overcome these challenges, we adopted a method used by (Lee et al. 2024) that transforms EHR data into a textual format via a structured template. This approach creates text "narratives" from the raw tabular data, which prevents the hallucinations typically associated with generative models (Hegselmann et al. 2024). This text-based data representation has shown to be effective in few-shot learning scenarios as demonstrated by (Hegselmann et al. 2023), and subsequent studies (Ono and Lee 2024) have reported competitive results when fine-tuned against traditional machine learning models.

Mathematically, this transformation can be expressed as a function $f : \mathcal{T} \to \mathcal{X}_{text}$, where $f(T_1, T_2, \ldots, T_n) = \mathbf{X}_{text}$. In this function, $T_1, T_2, \ldots, T_n$ represent the individual tables in the EHR, such as medications and diagnoses, and $\mathbf{X}_{text}$ is the resultant unified textual representation of a patient's data. The mapping for each table $f(T_i)$ is defined as:

$$f(T_i) = \sum_{k=1}^{m} g(x_k)$$

Here, $x_k \in T_i$ denotes the individual entries (e.g., lab values, diagnosis codes), and $g(x_k)$ converts each entry into a textual description. This transformation ensures that all tables can be aligned with a unique patient identifier and visit identifier, thus integrating various visits into a coherent structured narrative. A comprehensive list of the tabular data fields used is detailed in Table 6.

Benchmarking

In this study, we benchmark state-of-the-art foundation models in biology and medicine to evaluate their performance on eight specific antibiotic prediction tasks, each defined as a binary classification problem to determine antibiotic efficacy. This benchmarking is necessary due to the rapid development of these models and claims by (Lee, Lee, and Chiang 2024) that foundation models are inadequately assessed. A summary of the foundation models under review is presented in Table 2.

Models are tested using their pre-trained parameters, which remain unchanged during inference. Predictions are generated by extracting the classification (CLS) token from the models and inputting it into a LightGBM model (Ke et al. 2017) for final prediction output. Performance is assessed using several key metrics: the Matthews Correlation Coefficient (MCC), the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), the Area Under the Precision-Recall Curve (PRC-AUC), and the F1-score. These metrics evaluate how effectively each model handles the nuances of biomedical text and patient data.

The MCC provides a balanced measure across classes in imbalanced datasets, preventing an overemphasis on the majority class. The ROC-AUC measures the model's ability to distinguish between classes, which is critical for identifying susceptible versus resistant outcomes. The PRC-AUC is important for its focus on precision and recall, crucial in biomedical contexts where the costs of false positives or negatives are significant. The F1-score, which combines precision and recall, serves as a practical performance measure when both error types are significant. To ensure robust statistical validity, we bootstrap and resample our test set 1,000 times to establish 95% confidence intervals for each metric.

Furthermore, we rank each model for each antibiotic based on these metrics to determine which performs best on average. This analysis helps us quantify the performance improvements relative to the top-performing model, providing insights into the practical implications of model advancements.

Table 2: Foundation Models evaluated in our Benchmarking Study. We take a wide range of foundation models found in biomedicine that were trained on EHR sequences, scientific text, discharge summaries, among many others. Some of these models were trained solely on the Masked language modeling objective while others were trained in more specialized tasks like named entity recognition.

Name/HuggingFace Model Card	Source	Name/HuggingFace Model Card	Source
pritamdeka/BioBert-PubMed200kRCT	(Deka, Jurek-Loughrey 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, Leskovec, and Liang 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, Yan, and Lu 2019)	StanfordAIMI/RadBERT	(Chambon, Cook, and Langlotz 2022)
medicalai/ClinicalBERT	(Wang et al. 2023)	allenai/scibert_scivocab_uncased	(Beltagy, Lo, and Cohan 2019)

These comprehensive metrics ensure a thorough evaluation of each model's discriminative ability and reliability in predicting antibiotic efficacy. We hypothesize that model performance will vary depending on the complexity and prevalence of the antibiotic-related tasks, with some models performing better in specific scenarios.

Evaluating Missed and Incorrect Prescriptions

Missed and incorrect predictions in antibiotic susceptibility models are significant concerns, particularly in clinical settings where errors can adversely affect patient outcomes. Clinicians require insights not only into the overall performance of language foundation models but also into their specific failures, notably in terms of false positives and false negatives. A false positive—where a model incorrectly predicts susceptibility to an antibiotic when the pathogen is actually resistant—may result in the wrong treatment, potentially prolonging infection and leading to ineffective therapy. Our study aims to mitigate these false positives to enhance AI-driven stewardship methods. Conversely, a false negative—where a model incorrectly predicts resistance when the pathogen is actually susceptible—could cause clinicians to bypass effective treatment options, potentially resorting to more aggressive or costly alternatives unnecessarily.

Investigating these types of errors is crucial for enhancing model reliability and building clinician trust. By meticulously analyzing and documenting instances of false positives and false negatives, researchers can pinpoint patterns or specific conditions under which the models falter, such as with certain antibiotics or within particular patient subgroups. This detailed insight is invaluable to clinicians as it aids their decision-making process by identifying potential error-prone areas in the models.

Topic Modeling and Concordance analysis provides Interpretability

To enhance the interpretability of our text-based EHR approach, we utilize BERTopic (Grootendorst 2022), a topic modeling technique that combines BERT embeddings with hierarchical clustering (HDBSCAN (McInnes et al. 2017)). Each patient's text-based EHR record, represented as a sequence of tokens $S_i = [s_{i1}, s_{i2}, \ldots, s_{im}]$, is embedded into a high-dimensional space $\mathbb{R}^d$ using BERT. This process results in a sequence of embeddings $E_i = [e_{i1}, e_{i2}, \ldots, e_{im}]$, where each e_{ij} belongs to $\mathbb{R}^d$.

The dimensionality of the embeddings is reduced using UMAP:

$$E_i' = \text{UMAP}(E_i)$$

where E_i' is mapped to a lower-dimensional space $\mathbb{R}^k$ and k represents the reduced dimensions. Subsequently, HDBSCAN is applied to these reduced embeddings to identify clusters, forming topic groups $T_1, T_2, \ldots, T_l$:

$$T_i = \text{HDBSCAN}(E')$$

These clusters provide insights into patient groups with similar medical histories or conditions, thereby offering a way to understand patterns in the model's internal representations. We then perform a concordance analysis taking two random exemplars to try to understand the decision-making processes of these foundation models. This analysis allows for the interpretability of how the model predicts the appropriateness of specific antibiotics for new patients, based on the internal representations and similarities identified within these topic clusters.

Results

Benchmarking Results

Table 3 displays the raw performance metrics—F1 score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and Precision-Recall Curve Area Under the Curve (PRC-AUC)—for various models in multiple antibiotic susceptibility prediction tasks. For each model, performance metrics are detailed for every task, providing an extensive overview of their effectiveness in different scenarios. The numbers in parentheses represent the 95% confidence intervals, which are derived from bootstrapping the test set 1,000 times. This method helps in estimating the uncertainty of the metrics and assessing the statistical significance of the results.

Table 4 aggregates these metrics by listing the average ranks of the models across all four performance indicators. The models are ranked from best to worst based on their average overall rank, with lower ranks indicating superior performance. Notably, BioClinicalBERT and PubMedBERT exhibit the lowest average ranks at 5.22 and 5.25, respectively, closely followed by SciBERT and BioMedRoBERTa. This ranking system provides a standardized measure for comparing the efficacy of different models across various

Table 3: Results for CLINDAMYCIN, ERYTHROMYCIN, GENTAMICIN, and LEVOFLOXACIN

Model/Task	CLINDAMYCIN				ERYTHROMYCIN				GENTAMICIN				LEVOFLOXACIN			
	F1	MCC	ROC-AUC	PRC-AUC	F1	MCC	ROC-AUC	PRC-AUC	F1	MCC	ROC-AUC	PRC-AUC	F1	MCC	ROC-AUC	PRC-AUC
BioclinicalBERT	0.777 (±0.031)	0.354 (±0.023)	**0.750 (±0.039)**	**0.788 (±0.047)**	0.661 (±0.044)	0.367 (±0.041)	0.754 (±0.040)	0.694 (±0.059)	0.978 (±0.009)	**0.397 (±0.037)**	0.684 (±0.122)	0.970 (±0.017)	0.817 (±0.027)	0.453 (±0.029)	0.801 (±0.036)	0.844 (±0.041)
MedBERT	0.780 (±0.029)	0.369 (±0.045)	0.742 (±0.040)	0.773 (±0.045)	0.655 (±0.042)	0.362 (±0.015)	0.755 (±0.038)	0.692 (±0.060)	0.977 (±0.008)	**0.397 (±0.020)**	**0.731 (±0.111)**	**0.973 (±0.017)**	0.810 (±0.029)	0.421 (±0.039)	0.779 (±0.037)	0.822 (±0.043)
DistilBERT	0.778 (±0.031)	0.371 (±0.034)	0.740 (±0.041)	0.768 (±0.055)	0.656 (±0.045)	0.353 (±0.048)	0.749 (±0.041)	0.682 (±0.067)	0.977 (±0.008)	**0.397 (±0.013)**	0.693 (±0.106)	0.972 (±0.017)	0.810 (±0.028)	0.428 (±0.027)	0.792 (±0.037)	0.841 (±0.039)
BioMegatron	0.775 (±0.030)	0.353 (±0.032)	0.726 (±0.040)	0.762 (±0.052)	0.659 (±0.040)	0.352 (±0.049)	0.764 (±0.038)	0.702 (±0.060)	0.978 (±0.009)	**0.397 (±0.035)**	0.663 (±0.114)	0.970 (±0.016)	0.810 (±0.030)	0.426 (±0.025)	0.791 (±0.038)	0.845 (±0.041)
BlueBERT	**0.783 (±0.031)**	**0.404 (±0.038)**	0.738 (±0.038)	0.768 (±0.049)	0.652 (±0.041)	0.346 (±0.017)	0.736 (±0.040)	0.670 (±0.063)	0.978 (±0.008)	0.367 (±0.042)	0.681 (±0.112)	0.967 (±0.020)	0.819 (±0.028)	0.466 (±0.012)	0.780 (±0.039)	0.829 (±0.040)
PubMedBERT	0.781 (±0.031)	0.375 (±0.031)	0.738 (±0.039)	0.773 (±0.052)	0.660 (±0.042)	0.369 (±0.026)	**0.770 (±0.038)**	**0.717 (±0.055)**	**0.979 (±0.008)**	0.393 (±0.043)	0.618 (±0.118)	0.962 (±0.020)	**0.829 (±0.029)**	0.499 (±0.014)	**0.816 (±0.035)**	**0.862 (±0.035)**
GatorTron	0.776 (±0.031)	0.364 (±0.024)	0.711 (±0.042)	0.753 (±0.049)	0.656 (±0.042)	0.357 (±0.046)	0.747 (±0.038)	0.675 (±0.061)	0.977 (±0.009)	**0.397 (±0.018)**	0.684 (±0.100)	0.970 (±0.018)	0.819 (±0.027)	0.457 (±0.050)	0.803 (±0.033)	0.855 (±0.035)
BioMedRoBERTa	0.775 (±0.030)	0.351 (±0.021)	0.731 (±0.041)	0.778 (±0.048)	0.667 (±0.042)	0.384 (±0.030)	0.768 (±0.039)	0.715 (±0.055)	0.978 (±0.008)	0.393 (±0.019)	0.629 (±0.127)	0.961 (±0.021)	0.826 (±0.026)	0.502 (±0.040)	0.800 (±0.035)	0.842 (±0.039)
SciBERT	0.780 (±0.030)	0.371 (±0.028)	0.740 (±0.038)	0.780 (±0.048)	0.648 (±0.041)	0.329 (±0.033)	0.743 (±0.040)	0.685 (±0.057)	0.978 (±0.009)	0.355 (±0.044)	0.704 (±0.116)	0.967 (±0.022)	0.817 (±0.029)	0.457 (±0.011)	0.803 (±0.035)	0.846 (±0.040)
BioLM	0.782 (±0.030)	0.369 (±0.016)	0.729 (±0.040)	0.764 (±0.049)	0.660 (±0.041)	0.364 (±0.047)	0.752 (±0.039)	0.682 (±0.062)	0.978 (±0.008)	0.367 (±0.036)	0.669 (±0.117)	0.966 (±0.020)	0.822 (±0.028)	0.468 (±0.022)	0.799 (±0.039)	0.828 (±0.047)
RadBERT	0.778 (±0.030)	0.364 (±0.025)	0.731 (±0.041)	0.772 (±0.048)	**0.670 (±0.043)**	**0.408 (±0.041)**	0.768 (±0.041)	0.708 (±0.058)	0.977 (±0.008)	**0.397 (±0.039)**	0.688 (±0.111)	0.969 (±0.019)	0.812 (±0.028)	0.437 (±0.031)	0.799 (±0.035)	0.842 (±0.043)
LinkBERT	**0.783 (±0.030)**	0.384 (±0.015)	0.739 (±0.038)	0.779 (±0.049)	0.645 (±0.041)	0.330 (±0.029)	0.731 (±0.040)	0.669 (±0.060)	0.978 (±0.008)	0.395 (±0.043)	0.699 (±0.119)	0.968 (±0.020)	0.828 (±0.027)	**0.510 (±0.012)**	0.800 (±0.038)	0.844 (±0.037)
ClinicalBERT	0.777 (±0.031)	0.354 (±0.038)	**0.750 (±0.039)**	**0.788 (±0.047)**	0.661 (±0.044)	0.367 (±0.026)	0.754 (±0.040)	0.694 (±0.059)	0.978 (±0.009)	**0.397 (±0.045)**	0.684 (±0.122)	0.970 (±0.017)	0.817 (±0.027)	0.453 (±0.013)	0.801 (±0.036)	0.844 (±0.041)
BioBERT	0.778 (±0.029)	0.373 (±0.024)	0.706 (±0.042)	0.737 (±0.054)	0.648 (±0.040)	0.331 (±0.046)	0.736 (±0.039)	0.660 (±0.064)	0.978 (±0.009)	0.381 (±0.018)	0.704 (±0.127)	0.964 (±0.023)	0.811 (±0.028)	0.432 (±0.050)	0.768 (±0.039)	0.809 (±0.046)

Model/Task	OXACILLIN				TETRACYCLINE				TRIMETHOPRIM/SULFA				VANCOMYCIN			
	F1	MCC	ROC-AUC	PRC-AUC	F1	MCC	ROC-AUC	PRC-AUC	F1	MCC	ROC-AUC	PRC-AUC	F1	MCC	ROC-AUC	PRC-AUC
BioclinicalBERT	0.791 (±0.030)	0.441 (±0.015)	0.779 (±0.038)	0.805 (±0.046)	0.904 (±0.019)	0.372 (±0.041)	0.673 (±0.062)	0.863 (±0.040)	0.910 (±0.017)	0.435 (±0.025)	0.706 (±0.058)	0.878 (±0.036)	0.742 (±0.035)	0.384 (±0.033)	0.771 (±0.036)	0.788 (±0.043)
MedBERT	0.785 (±0.031)	0.413 (±0.028)	0.771 (±0.035)	0.797 (±0.043)	0.904 (±0.019)	0.372 (±0.047)	0.646 (±0.061)	0.848 (±0.043)	0.910 (±0.018)	0.415 (±0.016)	0.723 (±0.054)	0.898 (±0.029)	0.731 (±0.036)	0.355 (±0.032)	0.759 (±0.037)	0.791 (±0.041)
DistilBERT	0.782 (±0.031)	0.409 (±0.013)	0.762 (±0.038)	0.767 (±0.055)	0.904 (±0.019)	**0.380** (±0.038)	0.669 (±0.060)	0.862 (±0.040)	0.908 (±0.019)	0.407 (±0.020)	0.708 (±0.057)	0.878 (±0.038)	0.725 (±0.033)	0.324 (±0.046)	0.756 (±0.036)	0.779 (±0.043)
BioMegatron	0.791 (±0.031)	0.433 (±0.014)	0.781 (±0.039)	0.803 (±0.048)	0.904 (±0.018)	0.369 (±0.045)	0.673 (±0.054)	0.872 (±0.036)	**0.911** (±0.017)	0.416 (±0.022)	0.723 (±0.053)	0.883 (±0.036)	0.734 (±0.036)	0.360 (±0.049)	0.759 (±0.038)	0.769 (±0.050)
BlueBERT	0.793 (±0.029)	0.441 (±0.036)	0.764 (±0.038)	0.787 (±0.046)	**0.905** (±0.019)	**0.380** (±0.011)	0.654 (±0.055)	0.861 (±0.039)	0.909 (±0.018)	0.410 (±0.042)	0.734 (±0.054)	0.894 (±0.036)	0.734 (±0.034)	0.376 (±0.030)	0.773 (±0.036)	0.796 (±0.042)
PubMedBERT	0.796 (±0.029)	0.454 (±0.019)	0.786 (±0.036)	0.815 (±0.042)	**0.905** (±0.019)	**0.380** (±0.043)	0.668 (±0.057)	0.867 (±0.037)	**0.911** (±0.018)	0.433 (±0.024)	0.695 (±0.056)	0.872 (±0.038)	0.729 (±0.033)	0.336 (±0.035)	0.772 (±0.036)	0.797 (±0.041)
GatorTron	0.788 (±0.030)	0.424 (±0.027)	0.770 (±0.038)	0.798 (±0.048)	0.904 (±0.018)	0.378 (±0.039)	0.634 (±0.058)	0.844 (±0.043)	**0.911** (±0.017)	**0.436** (±0.012)	0.710 (±0.055)	0.877 (±0.038)	0.731 (±0.036)	0.356 (±0.044)	0.756 (±0.037)	0.785 (±0.041)
BioMedRoBERTa	**0.797** (±0.029)	**0.465** (±0.021)	**0.800** (±0.033)	**0.824** (±0.042)	0.904 (±0.018)	0.372 (±0.040)	0.661 (±0.058)	0.866 (±0.037)	0.910 (±0.019)	0.418 (±0.018)	**0.751** (±0.049)	**0.909** (±0.029)	0.738 (±0.035)	0.399 (±0.048)	**0.787** (±0.037)	**0.808** (±0.045)
SciBERT	0.791 (±0.030)	0.443 (±0.029)	0.785 (±0.037)	0.803 (±0.047)	**0.905** (±0.019)	**0.380** (±0.031)	**0.676** (±0.060)	**0.876** (±0.033)	0.910 (±0.018)	0.424 (±0.017)	**0.751** (±0.051)	**0.900** (±0.033)	**0.750** (±0.035)	**0.417** (±0.034)	0.783 (±0.036)	0.802 (±0.040)
BioLM	**0.797** (±0.030)	0.449 (±0.023)	0.785 (±0.035)	0.802 (±0.047)	0.904 (±0.019)	0.372 (±0.050)	0.666 (±0.060)	0.859 (±0.040)	0.910 (±0.018)	0.424 (±0.013)	0.717 (±0.055)	0.887 (±0.035)	0.731 (±0.033)	0.343 (±0.037)	0.776 (±0.037)	0.797 (±0.045)
RadBERT	0.790 (±0.029)	0.429 (±0.026)	0.778 (±0.037)	0.805 (±0.045)	0.904 (±0.018)	0.371 (±0.045)	0.663 (±0.054)	0.873 (±0.036)	0.908 (±0.018)	0.410 (±0.020)	0.700 (±0.059)	0.884 (±0.035)	0.725 (±0.033)	0.334 (±0.020)	0.749 (±0.037)	0.776 (±0.044)
LinkBERT	0.792 (±0.031)	0.448 (±0.047)	0.773 (±0.039)	0.785 (±0.052)	0.904 (±0.018)	0.372 (±0.015)	0.670 (±0.061)	0.863 (±0.040)	**0.911** (±0.018)	0.424 (±0.033)	0.719 (±0.056)	0.892 (±0.032)	0.735 (±0.035)	0.382 (±0.019)	0.777 (±0.036)	0.803 (±0.041)
ClinicalBERT	0.791 (±0.030)	0.441 (±0.049)	0.779 (±0.038)	0.805 (±0.046)	0.904 (±0.019)	0.372 (±0.025)	0.673 (±0.062)	0.863 (±0.040)	0.910 (±0.017)	0.435 (±0.041)	0.706 (±0.058)	0.878 (±0.036)	0.742 (±0.035)	0.384 (±0.028)	0.771 (±0.036)	0.788 (±0.043)
BioBERT	0.788 (±0.030)	0.420 (±0.038)	0.774 (±0.037)	0.786 (±0.052)	0.904 (±0.018)	0.372 (±0.014)	0.637 (±0.061)	0.849 (±0.039)	0.908 (±0.018)	0.410 (±0.046)	0.722 (±0.056)	0.891 (±0.036)	0.731 (±0.034)	0.345 (±0.012)	0.758 (±0.037)	0.770 (±0.047)

Table 4: Average Ranks of Models Across Metrics

Model	F1	MCC	PRC-AUC	ROC-AUC	Avg. Rank
BioclinicalBERT	4.88	5.25	5.25	5.50	**5.22**
PubMedBERT	3.75	4.75	5.88	6.63	**5.25**
SciBERT	6.50	6.88	4.38	3.63	**5.34**
BioMedRoBERTa	5.63	5.88	4.63	5.88	**5.50**
LinkBERT	6.50	6.00	7.50	6.88	**6.72**
ClinicalBERT	8.50	6.88	6.25	6.50	**7.03**
BlueBERT	5.50	6.63	9.25	9.13	**7.62**
BioLM	6.38	7.50	9.13	7.63	**7.66**
BioMegatron	7.25	9.63	7.38	7.63	**7.97**
MedBERT	8.75	9.00	7.50	7.50	**8.19**
RadBERT	10.13	9.38	6.75	8.75	**8.75**
GatorTron	8.88	6.88	9.63	10.13	**8.88**
DistilBERT	10.63	9.75	9.38	9.00	**9.69**
BioBERT	11.75	10.63	12.13	10.25	**11.19**

Table 5: Percentage of Missed and Incorrect Prescriptions per Model

Model	Incorrect Prescriptions (%)	Missed Prescriptions (%)
DistilBERT	10.44%	1.60%
BioMegatron	10.91%	1.38%
MedBERT	11.61%	1.08%
BlueBERT	10.45%	1.61%
BioBERT	11.67%	1.12%
PubMedBERT	10.32%	1.51%
BioMedRoBERTa	9.32%	2.15%
ClinicalBERT	10.93%	1.23%
BioClinicalBERT	9.92%	1.64%
SciBERT	10.65%	1.41%
BioLM	10.93%	1.32%
RadBERT	10.36%	1.62%
LinkBERT	10.25%	1.73%
Gatortron	10.60%	1.48%

metrics and tasks, focusing on ranks rather than absolute scores to facilitate a consistent assessment of relative model performance. Given that these models are all derived from similar sources with different pre-training or other techniques, we believe these results fall in line with what we expected and is known in the literature (Lee, Lee, and Chiang 2024).

Missed and Improper Prescriptions

Table 5 presents data on the percentage of incorrect and missed prescriptions by various language models across all antibiotic prescription tasks. Incorrect prescriptions, or false positives, occur when a model recommends an inappropriate antibiotic, and missed prescriptions, or false negatives, occur when a model fails to recommend a necessary antibiotic. These errors are clinically significant because incorrect prescriptions can promote antibiotic resistance, while missed prescriptions can result in inadequate treatment of bacterial infections. We totaled the number of false positives and false negatives across all antibiotic tasks then divided by eight times the test sample size (Each sample across all antibiotic tasks are treated independently) to obtain percentages displayed in Table 6.

Among the evaluated models, BioMedRoBERTa has the lowest percentage of incorrect prescriptions at 9.32%, but it also has the highest percentage of missed prescriptions at 2.15%. Conversely, MedBERT exhibits a lower rate of missed prescriptions at 1.08% but a higher rate of incorrect prescriptions at 11.61%. PubMedBERT and BioClinicalBERT strike a balance, with incorrect prescription rates of 10.32% and 9.92%, respectively, and missed prescription rates between 1.5-1.6%.

The variation in these error rates across models underscores the trade-offs involved in minimizing incorrect versus missed prescriptions. From an antibiotic stewardship perspective, these results indicate that while the models are effective to some extent, there is still considerable room for improvement, especially in reducing the high rates of false positives. Managing incorrect prescriptions is crucial for effective treatment and combating antibiotic resistance.

Topic Modeling Results

Figure 2 displays a similarity matrix resulting from hierarchical topic modeling performed on electronic health records (EHR) text data using BERTopic (Grootendorst 2022). This technique groups similar topics together based on the semantic content of the EHR entries, which, upon further analysis, predominantly involved patient diagnoses and prescriptions. Utilizing the text-based structure of the EHR, this method revealed underlying patterns in clinical

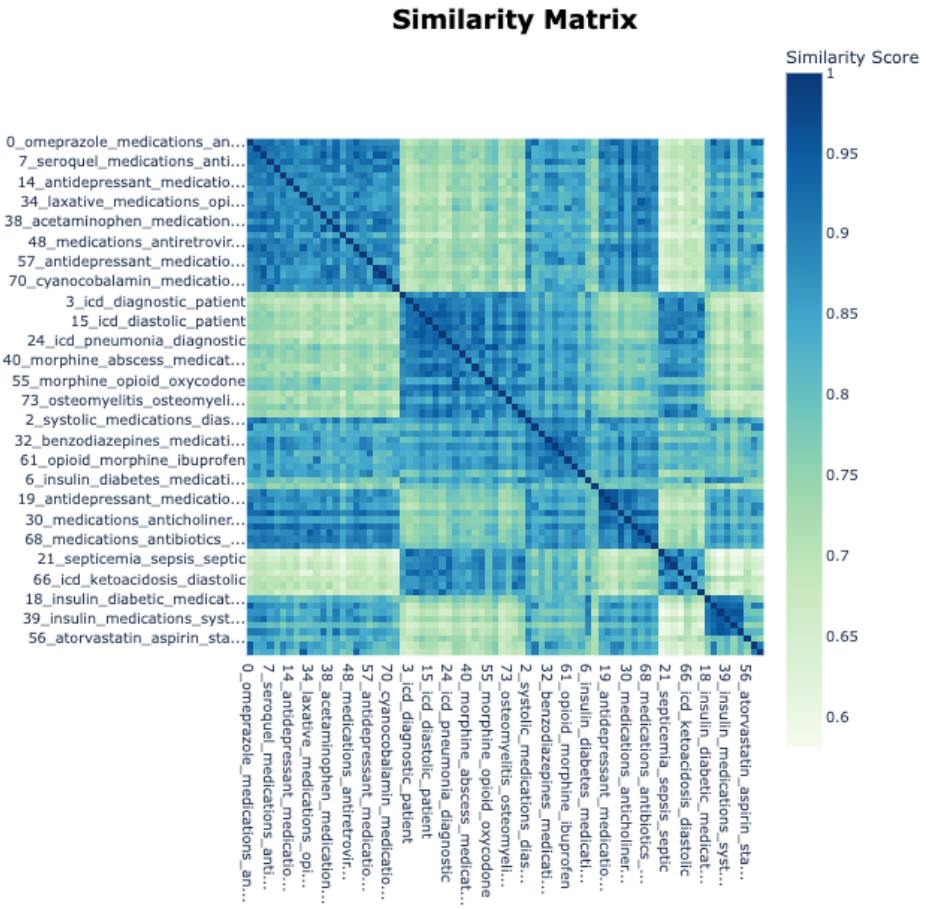

Figure 2: Similarity matrix showing the relationships between various clinical features or patient groups, with darker colors indicating higher similarity scores. Clusters of patients with similar characteristics are represented by dense squares. Topic modeling was applied to identify these clusters, and concordance analysis confirmed consistent model predictions for patients within the same group, suggesting coherent decision-making based on shared clinical profiles.

language, facilitating a thorough exploration of the relationships between various medical conditions and treatments.

To enhance our understanding of how models make decisions regarding antibiotic prescriptions, we performed a concordance analysis and sampled two random entries from each cluster identified by BERTopic. This sampling strategy added a layer of interpretability by allowing us to explore the nuances of model behavior within specific medical contexts. Our analysis identified distinct topic clusters that encompass key healthcare areas such as sepsis, diabetes, stomach acid disorders, anxiety, pain management, respiratory conditions, and the use of antidepressants.

The topic clusters provided insights into the effectiveness of antibiotics across different medical conditions, shedding light on how models categorize decisions. For example, in the sepsis cluster, a concordance study of two samples from the same group indicated that common antibiotics often failed, highlighting the critical need for accurate and timely antibiotic selection in sepsis treatment to avert severe outcomes like septic shock. Similarly, in clusters associated with diabetes and respiratory conditions, challenges were evident; for instance, infections in diabetic patients were often more difficult to manage, with certain antibiotics proving less effective due to underlying metabolic complications.

Discussion

High Accuracy but Over-Prediction of Antibiotics

The results show that foundation models such as BioClinicalBERT, PubMedBERT, and SciBERT consistently achieve high accuracy in antibiotic prediction tasks, with correct prescription rates averaging between 86% and 88% across various tasks. These outcomes highlight the capability of foundation models to enhance clinical decision-making by effectively automating antibiotic selection. Despite these strong performances, the models exhibit a tendency to overpredict the positive label, that is, recommending antibiotics more frequently than necessary. This propensity for overprescription, reflected in higher rates of incorrect prescriptions (false positives), presents a challenge in the context of antibiotic stewardship.

Antibiotic stewardship programs are designed to mini-

mize unnecessary antibiotic use to combat the rising threat of antibiotic resistance. False positives, where an antibiotic is unnecessarily recommended, are problematic. Although the models demonstrate high accuracy in identifying appropriate cases for antibiotic use, their bias towards over-predicting positive cases could contribute to antibiotic overuse in clinical settings, potentially fostering resistance and other negative outcomes. Future efforts aim to refine these models to better balance sensitivity and specificity, thereby reducing the incidence of false positives without diminishing their ability in detecting true positive cases.

Trade-offs Between Incorrect and Missed Prescriptions

The results reveal a critical trade-off between incorrect (false positive) and missed prescriptions (false negative). For instance, BioMedRoBERTa achieves the lowest rate of incorrect prescriptions at 9.32%, but this comes at the cost of the highest rate of missed prescriptions at 2.15%. Conversely, models like MedBERT and PubMedBERT exhibit lower rates of missed prescriptions but have higher rates of incorrect prescriptions. This trade-off highlights the complexity of optimizing model performance in clinical settings, where both over- and under-prescription can have significant consequences.

From a clinical perspective, reducing missed prescriptions is essential to ensure that patients receive appropriate and timely antibiotic treatments, particularly in severe conditions such as sepsis. However, minimizing incorrect prescriptions is equally important to avoid contributing to antibiotic resistance. Future research should focus on balancing these two types of errors, enhancing the models' capacity to avoid both over-prescription and under-prescription.

Interpretability and Model Decision Insights

The topic modeling results yield important insights into how foundation models categorize clinical data and approach clinical outcome tasks, despite their "black box" nature as deep learning systems. Utilizing BERTopic on EHR text data, we identified clusters or topics associated with conditions such as sepsis, diabetes, and respiratory diseases. These clusters reveal the internal representations that foundation models capture.

These findings not only improve the interpretability of model decisions but also pinpoint areas for further refinement of foundation models. For instance, incorporating clinical guidelines or more comprehensive patient data could enhance the models' ability to differentiate between cases requiring distinct antibiotic strategies. Future research might explore using similarity-based searches within these topic clusters to tailor antibiotic selection more precisely, thereby enhancing both prediction accuracy and clinical applicability.

In summary, while foundation models demonstrate potential in antibiotic prescription tasks, addressing their propensity to over-predict and achieving an optimal balance between incorrect and missed prescriptions remain critical for their practical application in clinical settings. Continued efforts to improve model interpretability and to incorporate more patient-specific data are essential for advancing both antibiotic stewardship and model performance.

Conclusion

This study proposed a framework to evaluated the performance of biomedical language models in predicting antibiotic susceptibility from electronic health records (EHRs), specifically focusing on eight crucial antibiotics used to treat infections caused by Staphylococcus aureus. Although the models generally predicted effective antibiotics accurately, they exhibited a tendency to over-predict susceptibility, resulting in a higher incidence of incorrect prescriptions. This tendency is particularly problematic within the framework of antibiotic stewardship, as unnecessary antibiotic usage can hasten the development of resistant bacterial strains.

The analysis highlighted a trade-off between incorrect and missed prescriptions among the evaluated models. Some models reduced incorrect prescriptions but increased missed prescriptions, illustrating the complexity involved in optimizing model performance for clinical use. Additionally, through topic modeling and interpretability analyses, it was observed that models frequently make similar decisions based on the representations of clinical concepts within their internal frameworks.

Collaboration between AI researchers, clinicians, and healthcare institutions is essential to develop models that are technically sound, clinically relevant, and ethically robust. Enhancing the capabilities of AI systems and addressing their current limitations are vital steps towards harnessing AI-driven decision support to improve antibiotic stewardship and address the global challenge of antibiotic resistance.

Limitations A significant limitation of this study is its reliance on data from a single source—the MIMIC-IV-ED dataset—which might not reflect the diversity of patient populations and clinical scenarios encountered in different healthcare settings. This limitation could affect the generalizability and applicability of the findings.

Future Works Future research should consider integrating more comprehensive patient data and clinical decision rules into the predictive models. A study on fairness could also be conducted to evaluate how much of the error was caused by patients in imbalanced groups (e.g. Asian or Black patients).

Another promising avenue is the use of decoder-based generative large language models (LLMs), especially those with extensive parameter counts (e.g., 70 billion parameters). These models have shown remarkable proficiency in understanding and generating complex language patterns, which could potentially be applied to capture detailed clinical nuances and patient-specific factors.

Acknowledgments We thank H.H. and Y.H. who were participants of the UCLA Bruins in Genomics (B.I.G) program for their contributions.

References

Adegoke, A. A.; Faleye, A. C.; Singh, G.; and Stenström, T. A. 2016. Antibiotic resistant superbugs: assessment of the interrelationship of occurrence in clinical settings and environmental niches. *Molecules*, 22(1): 29.

Al Kuwaiti, A.; Nazer, K.; Al-Reedy, A.; Al-Shehri, S.; Al-Muhanna, A.; Subbarayalu, A. V.; Al Muhanna, D.; and Al-Muhanna, F. A. 2023. A review of the role of artificial intelligence in healthcare. *Journal of personalized medicine*, 13(6): 951.

Alpert, P. T. 2017. Superbugs: antibiotic resistance is becoming a major public health concern. *Home Health Care Management & Practice*, 29(2): 130–133.

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Antal, M.; Marosi, M.; Nagy, T.; Juhász, G.; and Antal, P. 2024. M-BioBERTa: Modular RoBERTa-based Model for Biobank-scale Unified Representations.

Ávila-Jiménez, J. L.; Cantón-Habas, V.; del Pilar Carrera-González, M.; Rich-Ruiz, M.; and Ventura, S. 2024. A deep learning model for Alzheimer's disease diagnosis based on patient clinical records. *Computers in Biology and Medicine*, 169: 107814.

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

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Chambon, P.; Cook, T. S.; and Langlotz, C. P. 2022. Improved fine-tuning of in-domain transformer model for inferring COVID-19 presence in multi-institutional radiology reports. *Journal of Digital Imaging*.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 787–795.

De Kraker, M. E.; Davey, P. G.; Grundmann, H.; and Group, B. S. 2011. Mortality and hospital stay associated with resistant Staphylococcus aureus and Escherichia coli bacteremia: estimating the burden of antibiotic resistance in Europe. *PLoS medicine*, 8(10): e1001104.

Deka, P.; Jurek-Loughrey, A.; et al. 2022. Evidence Extraction to Validate Medical Claims in Fake News Detection. In *International Conference on Health Information Science*, 3–15. Springer.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Farnoud, A.; Ohnmacht, A. J.; Meinel, M.; and Menden, M. P. 2022. Can artificial intelligence accelerate preclinical drug discovery and precision medicine? *Expert Opinion on Drug Discovery*, 17(7): 661–665.

Gandra, S.; Tseng, K. K.; Arora, A.; Bhowmik, B.; Robinson, M. L.; Panigrahi, B.; Laxminarayan, R.; and Klein, E. Y. 2019. The mortality burden of multidrug-resistant pathogens in India: a retrospective, observational study. *Clinical Infectious Diseases*, 69(4): 563–570.

Gao, Y.; Shang, Q.; Li, W.; Guo, W.; Stojadinovic, A.; Mannion, C.; Man, Y.-g.; and Chen, T. 2020. Antibiotics for cancer treatment: A double-edged sword. *Journal of Cancer*, 11(17): 5135.

Golkar, Z.; Bagasra, O.; and Pace, D. G. 2014. Bacteriophage therapy: a potential solution for the antibiotic resistance crisis. *The Journal of Infection in Developing Countries*, 8(02): 129–136.

Gould, I. M.; and Bal, A. M. 2013. New antibiotic agents in the pipeline and how they can help overcome microbial resistance. *Virulence*, 4(2): 185–191.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. .

Gupta, M.; Gallamoza, B.; Cutrona, N.; Dhakal, P.; Poulain, R.; and Beheshti, R. 2022. An extensive data processing pipeline for mimic-iv. In *Machine Learning for Health*, 311–325. PMLR.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of ACL*.

Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 1–10.

Hager, P.; Jungmann, F.; and Rueckert, D. ???? MIMIC-IV-Ext Clinical Decision Making: A MIMIC-IV Derived Dataset for Evaluation of Large Language Models on the Task of Clinical Decision Making for Abdominal Pathologies.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, 5549–5581. PMLR.

Hegselmann, S.; Shen, S. Z.; Gierse, F.; Agrawal, M.; Sontag, D.; and Jiang, X. 2024. A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models. *arXiv preprint arXiv:2402.15422*.

Hindler, J. F.; and Munro, S. 2010. Antimicrobial susceptibility testing. *Clinical microbiology procedures handbook*, 5–0.

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

Idowu, E. A. A.; Teo, J.; Salih, S.; Valverde, J.; and Yeung, J. A. 2023. Streams, rivers and data lakes: an introduction to understanding modern electronic healthcare records. *Clinical Medicine*, 23(4): 409–413.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Jones, D. J.; Bunn, F.; and Bell-Syer, S. V. 2014. Prophylactic antibiotics to prevent surgical site infection after breast cancer surgery. *Cochrane Database of Systematic Reviews*, (3).

Jorgensen, J. H.; and Turnidge, J. D. 2015. Susceptibility test methods: dilution and disk diffusion methods. *Manual of clinical microbiology*, 1253–1273.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Khalighi, S.; Reddy, K.; Midya, A.; Pandav, K. B.; Madabhushi, A.; and Abedalthagafi, M. 2024. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ Precision Oncology*, 8(1): 80.

Kohli, R.; and Tan, S. S.-L. 2016. Electronic health records. *Mis Quarterly*, 40(3): 553–574.

Kumar, R. P.; Sivan, V.; Bachir, H.; Sarwar, S. A.; Ruzicka, F.; O'Malley, G. R.; Lobo, P.; Morales, I. C.; Cassimatis, N. D.; Hundal, J. S.; et al. 2024. Can Artificial Intelligence Mitigate Missed Diagnoses by Generating Differential Diagnoses for Neurosurgeons? *World Neurosurgery*.

Lalitha, M.; et al. 2004. Manual on antimicrobial susceptibility testing. *Performance standards for antimicrobial testing: Twelfth Informational Supplement*, 56238: 454–456.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Lee, S. A.; Brokowski, T.; and Chiang, J. N. 2024. Enhancing Antibiotic Stewardship using a Natural Language Approach for Better Feature Representation. *arXiv preprint arXiv:2405.20419*.

Lee, S. A.; Jain, S.; Chen, A.; Biswas, A.; Fang, J.; Rudas, A.; and Chiang, J. N. 2024. Multimodal clinical pseudonotes for emergency department prediction tasks using multiple embedding model for ehr (meme). *arXiv preprint arXiv:2402.00160*.

Lee, S. A.; Lee, J.; and Chiang, J. N. 2024. FEET: A Framework for Evaluating Embedding Techniques. arXiv:2411.01322.

Lee, S. A.; and Lindsey, T. 2024. Do Large Language Models understand Medical Codes? *arXiv preprint arXiv:2403.10822*.

Li, L.; Zhou, J.; Gao, Z.; Hua, W.; Fan, L.; Yu, H.; Hagen, L.; Zhang, Y.; Assimes, T. L.; Hemphill, L.; et al. 2024. A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs). *arXiv preprint arXiv:2405.03066*.

Li, M. M.; Huang, K.; and Zitnik, M. 2022. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6(12): 1353–1369.

Li, Y.; Rao, S.; Solares, J. R. A.; Hassaine, A.; Ramakrishnan, R.; Canoy, D.; Zhu, Y.; Rahimi, K.; and Salimi-Khorshidi, G. 2020. BEHRT: transformer for electronic health records. *Scientific reports*, 10(1): 7155.

Lu, L.; Jin, P.; Pang, G.; Zhang, Z.; and Karniadakis, G. E. 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3): 218–229.

Lu, Y.; and Lu, J. 2020. A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in neural information processing systems*, 33: 3094–3105.

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

Ma, M. D.; Ye, C.; Yan, Y.; Wang, X.; Ping, P.; Chang, T.; and Wang, W. 2024. CliBench: Multifaceted Evaluation of Large Language Models in Clinical Decisions on Diagnoses, Procedures, Lab Tests Orders and Prescriptions.

Magrabi, F.; Ammenwerth, E.; McNair, J. B.; De Keizer, N. F.; Hyppönen, H.; Nykänen, P.; Rigby, M.; Scott, P. J.; Vehko, T.; Wong, Z. S.-Y.; et al. 2019. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. *Yearbook of medical informatics*, 28(01): 128–134.

McInnes, L.; Healy, J.; Astels, S.; et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11): 205.

Miotto, R.; Wang, F.; Wang, S.; Jiang, X.; and Dudley, J. T. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6): 1236–1246.

Mullowney, M. W.; Duncan, K. R.; Elsayed, S. S.; Garg, N.; van der Hooft, J. J.; Martin, N. I.; Meijer, D.; Terlouw, B. R.; Biermann, F.; Blin, K.; et al. 2023. Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery*, 22(11): 895–916.

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

Nick, T. G.; and Campbell, K. M. 2007. Logistic regression. *Topics in biostatistics*, 273–301.

Nielsen, M. 2016. A visual proof that neural nets can compute any function. *URL: http://neuralnetworksanddeeplearning. com/chap4. html*.

Ono, K.; and Lee, S. A. 2024. Text Serialization and Their Relationship with the Conventional Paradigms of Tabular Machine Learning. *arXiv preprint arXiv:2406.13846*.

Peng, Y.; Yan, S.; and Lu, Z. 2019. 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)*, 58–65.

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

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; and Zhi, D. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1): 86.

Read, A. F.; and Woods, R. J. 2014. Antibiotic resistance management. *Evolution, medicine, and public health*, 2014(1): 147.

Reller, L. B.; Weinstein, M.; Jorgensen, J. H.; and Ferraro, M. J. 2009. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. *Clinical infectious diseases*, 49(11): 1749–1755.

Rupp, M.; Peter, O.; and Pattipaka, T. 2023. Exbehrt: Extended transformer for electronic health records. In *International Workshop on Trustworthy Machine Learning for Healthcare*, 73–84. Springer.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sengupta, S.; Chattopadhyay, M. K.; and Grossart, H.-P. 2013. The multifaceted roles of antibiotics and antibiotic resistance in nature. *Frontiers in microbiology*, 4: 47.

Shin, H.-C.; Zhang, Y.; Bakhturina, E.; Puri, R.; Patwary, M.; Shoeybi, M.; and Mani, R. 2020. BioMegatron: Larger Biomedical Domain Language Model. arXiv:2010.06060.

Shortliffe, E. H.; and Sepúlveda, M. J. 2018. Clinical decision support in the era of artificial intelligence. *Jama*, 320(21): 2199–2200.

Song, X.; Salcianu, A.; Song, Y.; Dopson, D.; and Zhou, D. 2020. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J. N.; Wu, Z.; and Ding, X. 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63: 101693.

Tenover, F. C.; and Gorwitz, R. J. 2006. The epidemiology of Staphylococcus infections. *Gram-positive pathogens*, 526–534.

Thornsberry, C.; and Mcdougal, L. K. 1983. Successful use of broth microdilution in susceptibility tests for methicillin-resistant (heteroresistant) staphylococci. *Journal of clinical microbiology*, 18(5): 1084–1091.

Vasantharajan, C.; Tun, K. Z.; Thi-Nga, H.; Jain, S.; Rong, T.; and Siong, C. E. 2022. 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)*, 1482–1488.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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

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

Wang, G.; Liu, X.; Ying, Z.; Yang, G.; Chen, Z.; Liu, Z.; Zhang, M.; Yan, H.; Lu, Y.; Gao, Y.; et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10): 2633–2642.

Warnke, P. H.; Lott, A. J.; Sherry, E.; Wiltfang, J.; and Podschun, R. 2013. The ongoing battle against multi-resistant strains: in-vitro inhibition of hospital-acquired MRSA, VRE, Pseudomonas, ESBL E. coli and Klebsiella species in the presence of plant-derived antiseptic oils. *Journal of Cranio-Maxillofacial Surgery*, 41(4): 321–326.

Wornow, M.; Thapa, R.; Steinberg, E.; Fries, J.; and Shah, N. 2023. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36: 67125–67137.

Xie, F.; Yuan, H.; Ning, Y.; Ong, M. E. H.; Feng, M.; Hsu, W.; Chakraborty, B.; and Liu, N. 2022. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of biomedical informatics*, 126: 103980.

Yang, X.; Chen, A.; PourNejatian, N.; Shin, H. C.; Smith, K. E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A. B.; Flores, M. G.; et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1): 194.

Yang, Z.; Batra, S. S.; Stremmel, J.; and Halperin, E. 2023a. Surpassing GPT-4 Medical Coding with a Two-Stage Approach. *arXiv preprint arXiv:2311.13735*.

Yang, Z.; Mitra, A.; Liu, W.; Berlowitz, D.; and Yu, H. 2023b. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1): 7857.

Yasunaga, M.; Leskovec, J.; and Liang, P. 2022. LinkBERT: Pretraining Language Models with Document Links. In *Association for Computational Linguistics (ACL)*.

Zhou, H.-Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; and Li, W. 2023. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature biomedical engineering*, 7(6): 743–755.

EHR Concepts Table

Table 6: Overview of Clinical Tables transformed from tabular to text in the MIMIC Emergency Department Visits Data.

Modality Name	Description
Arrival Information	Records patient demographics, time of arrival, and mode of arrival (e.g., ambulance, walk-in).
Triage Information	Documents vital signs, severity of condition using scales like ESI, and initial chief complaints upon arrival.
Medication Reconciliation	Details previous and current medications the patient is taking, including dosages and frequency.
Patient Vitals	Ongoing measurements throughout the ED visit including heart rate, blood pressure, temperature, etc.
Diagnosis Codes	ICD-9/10 codes used to classify and record diagnoses during the visit.
Pyxis Information	Information on medications administered during the ED stay via the Pyxis system, including timing and dosage.

Full Results

Table 7: False Positives and False Negatives per Antibiotic per Model (with Percentages) - Part 1

Model	**Clindamycin**		**Erythromycin**		**Gentamicin**		**Levofloxacin**	
	FP	FN	FP	FN	FP	FN	FP	FN
DistilBERT	168 (14.04%)	15 (1.25%)	156 (13.03%)	44 (3.68%)	23 (1.92%)	4 (0.33%)	122 (10.19%)	34 (2.84%)
BioMegatron	177 (14.79%)	16 (1.34%)	176 (14.71%)	39 (3.26%)	23 (1.92%)	4 (0.33%)	146 (12.20%)	18 (1.50%)
MedBERT	149 (12.45%)	29 (2.42%)	240 (20.05%)	17 (1.42%)	23 (1.92%)	4 (0.33%)	145 (12.11%)	14 (1.17%)
BlueBERT	142 (11.86%)	32 (2.67%)	221 (18.46%)	18 (1.50%)	25 (2.09%)	2 (0.17%)	129 (10.77%)	23 (1.92%)
BioBERT	181 (15.12%)	12 (1.00%)	214 (17.88%)	25 (2.09%)	24 (2.00%)	2 (0.17%)	144 (12.03%)	21 (1.75%)
PubMedBERT	164 (13.70%)	21 (1.75%)	139 (11.61%)	57 (4.76%)	26 (2.17%)	0 (0.00%)	114 (9.52%)	28 (2.34%)
BioMedRoBERTa	179 (14.96%)	15 (1.25%)	160 (13.36%)	41 (3.43%)	26 (2.17%)	0 (0.00%)	101 (8.44%)	39 (3.26%)
ClinicalBERT	178 (14.88%)	15 (1.25%)	197 (16.46%)	25 (2.09%)	23 (1.92%)	4 (0.33%)	150 (12.53%)	10 (0.84%)
BioClinicalBERT	174 (14.53%)	16 (1.34%)	134 (11.19%)	49 (4.09%)	24 (2.00%)	3 (0.25%)	141 (11.78%)	13 (1.09%)
SciBERT	178 (14.88%)	12 (1.00%)	211 (17.64%)	26 (2.17%)	26 (2.17%)	1 (0.08%)	135 (11.28%)	20 (1.67%)
BioLM	151 (12.61%)	28 (2.34%)	204 (17.04%)	22 (1.84%)	25 (2.09%)	2 (0.17%)	111 (9.27%)	35 (2.92%)
RadBERT	174 (14.53%)	16 (1.34%)	111 (9.28%)	66 (5.51%)	23 (1.92%)	4 (0.33%)	147 (12.28%)	15 (1.25%)
LinkBERT	170 (14.20%)	15 (1.25%)	237 (19.80%)	14 (1.17%)	25 (2.09%)	1 (0.08%)	103 (8.60%)	35 (2.92%)
Gatortron	168 (14.04%)	19 (1.56%)	185 (15.43%)	34 (2.85%)	24 (2.00%)	2 (0.17%)	130 (10.86%)	23 (1.92%)

Table 8: False Positives and False Negatives per Antibiotic per Model (with Percentages) - Part 2

Model	**Oxacillin**		**Tetracycline**		**Trimethoprim/Sulfa**		**Vancomycin**	
	FP	FN	FP	FN	FP	FN	FP	FN
DistilBERT	160 (13.36%)	18 (1.50%)	91 (7.60%)	9 (0.75%)	88 (7.35%)	8 (0.67%)	192 (16.04%)	21 (1.75%)
BioMegatron	149 (12.45%)	22 (1.84%)	94 (7.86%)	7 (0.58%)	86 (7.19%)	9 (0.75%)	194 (16.20%)	17 (1.42%)
MedBERT	165 (13.78%)	13 (1.09%)	91 (7.60%)	10 (0.84%)	86 (7.19%)	8 (0.67%)	213 (17.79%)	8 (0.67%)
BlueBERT	168 (14.04%)	8 (0.67%)	91 (7.60%)	9 (0.75%)	85 (7.10%)	11 (0.92%)	140 (11.70%)	51 (4.26%)
BioBERT	177 (14.79%)	9 (0.75%)	91 (7.60%)	10 (0.84%)	80 (6.68%)	15 (1.25%)	206 (17.21%)	13 (1.09%)
PubMedBERT	147 (12.28%)	19 (1.59%)	93 (7.77%)	7 (0.58%)	89 (7.44%)	4 (0.33%)	216 (18.05%)	9 (0.75%)
BioMedRoBERTa	122 (10.19%)	36 (3.01%)	91 (7.60%)	10 (0.84%)	84 (7.02%)	11 (0.92%)	129 (10.77%)	54 (4.51%)
ClinicalBERT	163 (13.62%)	11 (0.92%)	91 (7.60%)	10 (0.84%)	82 (6.85%)	11 (0.92%)	163 (13.62%)	32 (2.67%)
BioClinicalBERT	139 (11.61%)	19 (1.59%)	93 (7.77%)	8 (0.67%)	85 (7.10%)	10 (0.84%)	160 (13.36%)	39 (3.26%)
SciBERT	132 (11.03%)	33 (2.76%)	91 (7.60%)	9 (0.75%)	86 (7.19%)	8 (0.67%)	161 (13.45%)	26 (2.17%)
BioLM	164 (13.70%)	12 (1.00%)	91 (7.60%)	10 (0.84%)	87 (7.27%)	7 (0.58%)	214 (17.87%)	10 (0.84%)
RadBERT	156 (13.03%)	18 (1.50%)	92 (7.68%)	9 (0.75%)	85 (7.10%)	11 (0.92%)	204 (17.04%)	16 (1.34%)
LinkBERT	133 (11.11%)	31 (2.59%)	91 (7.60%)	10 (0.84%)	86 (7.19%)	8 (0.67%)	137 (11.44%)	52 (4.34%)
Gatortron	152 (12.70%)	19 (1.59%)	92 (7.68%)	9 (0.75%)	85 (7.10%)	9 (0.75%)	179 (14.96%)	27 (2.26%)