

Integrating Radiology Report Text and Structured Clinical Features for Predicting One- Year Mortality in Osteomyelitis: A Retrospective Multimodal Study

Yuhang Jin

Shandong University of Traditional Chinese Medicine

Zhipeng Fan

960th Hospital of PLA

Xuesong Yan

960th Hospital of PLA

Jiawei Li

960th Hospital of PLA

Zhuohang Wu

960th Hospital of PLA

Yongxian Zhang

Zhangyongxian513@126.com

960th Hospital of PLA

Research Article

Keywords: osteomyelitis, radiology reports, machine learning, mortality prediction, MIMIC-IV, multimodal modelling

Posted Date: August 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10431408/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Osteomyelitis remains a serious musculoskeletal infection with substantial clinical heterogeneity and a high risk of adverse outcomes. Conventional risk assessment often depends on a narrow set of demographic and laboratory variables, which may not capture the full spectrum of information contained in radiology reports and longitudinal electronic health record data. We therefore evaluated whether a multimodal machine learning framework combining unstructured radiology report text with structured clinical features could improve prediction of one-year mortality in patients with osteomyelitis.

Methods

We conducted a retrospective study using the MIMIC-IV database and linked radiology report information. Patients with osteomyelitis-related radiology reports were included, and one-year mortality was defined as the primary outcome. Structured predictors included demographics, comorbidities, vital signs, and laboratory measurements. Unstructured predictors were obtained from radiology reports using sentence-transformer embeddings. We compared logistic regression, random forest, LightGBM, XGBoost, and a stacking ensemble. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), calibration, accuracy, and F1 score on an independent test set, with 95% confidence intervals estimated by bootstrap resampling.

Results

The final analytical dataset comprised 97,594 records after preprocessing and harmonization. The stacking ensemble achieved the highest discrimination, with an AUC of 0.795 (95% CI 0.787–0.803), accuracy of 0.814, and F1 score of 0.414. The model demonstrated clinically meaningful discrimination and acceptable calibration. The integration of radiology report embeddings with structured clinical variables yielded informative predictive performance beyond conventional approaches.

Conclusions

A multimodal strategy that combines radiology report text and structured clinical data can provide clinically useful prediction of one-year mortality in osteomyelitis. These findings support the broader application of artificial intelligence methods for risk stratification in musculoskeletal infection and may inform individualized clinical monitoring and treatment planning.

Introduction

Osteomyelitis remains a clinically important and therapeutically challenging condition because it combines infection, inflammation, tissue destruction, and marked biological heterogeneity. Its presentation ranges from localized infection to severe systemic illness, and outcomes are shaped not only by the microbial process itself but also by age, comorbidity burden, organ dysfunction, and the

presence of complications such as abscess formation, sepsis, or chronic tissue destruction. In routine practice, clinicians synthesize history, examination findings, laboratory abnormalities, and imaging interpretation to estimate prognosis(1–3). However, such synthesis is often subjective and incomplete, especially in high-volume settings where clinical information is fragmented across structured and unstructured sources(4–6).

The rapid expansion of electronic health record systems has created new opportunities for data-driven prediction in medicine. Machine learning approaches have shown encouraging performance for outcomes such as mortality, readmission, and acute organ injury using structured clinical data(7, 8). Nevertheless, many existing models remain limited by their dependence on coded variables and therefore fail to capture clinically relevant signals embedded in narrative documentation(8–10). Radiology reports are particularly informative in this context. They contain nuanced descriptions of anatomical extent, chronicity, severity, diagnostic uncertainty, and progression, all of which may reflect disease burden and may not be adequately represented in structured fields(11, 12).

Recent advances in natural language processing have made it feasible to transform unstructured clinical text into clinically informative features. Transformer-based embedding methods, including sentence-level representations, have demonstrated strong performance in biomedical text mining and clinical prediction tasks(13–15). In parallel, multimodal modelling has emerged as a promising direction in medicine, combining heterogeneous data sources such as tabular clinical variables, imaging, and text to improve prediction(15–17). These developments are especially relevant to osteomyelitis, in which diagnosis and monitoring depend heavily on imaging interpretation and in which the language of radiology reports may contain prognostic information beyond coded diagnoses and laboratory abnormalities(18, 19).

We therefore developed and evaluated a multimodal machine learning framework to predict one-year mortality in patients with osteomyelitis by integrating structured clinical features and radiology report text. We hypothesized that combining these two complementary data modalities would yield better predictive performance than structured variables alone and would provide a clinically meaningful foundation for risk stratification in musculoskeletal infection.

Methods

Study design and data source

This retrospective study used the MIMIC-IV database, a large de-identified clinical repository that has been widely used for biomedical informatics research(19). We linked available radiology report text to structured clinical variables and defined the primary outcome as death within one year of the index clinical encounter(5). The study used de-identified public data and did not require additional ethics approval beyond the governance requirements of the source database.

Cohort definition

Patients were identified through osteomyelitis-related radiology reports and associated clinical records. The dataset was harmonized to remove duplicate or inconsistent entries and to ensure that the analytical structure reflected patient-level outcomes rather than repeated observations. Records with missing key variables or ambiguous outcome information were excluded according to predefined quality-control rules. The final analytic dataset was organized for patient-level modelling to minimize leakage from repeated measurements.

The patient selection process is illustrated in **Fig. 1**. A total of 6,127 patients with an osteomyelitis diagnosis were initially identified from the MIMIC-IV database. After linking with the radiology reports module, 97,594 reports were retrieved. Following patient-level deduplication, 2,934 patients with complete data were included in the final analytical cohort, which was subsequently divided into training (70%, $n = 2,053$), validation (15%, $n = 441$), and test (15%, $n = 440$) sets.

Figure 1. Patient flow diagram.

[Insert Fig. 1 here]

This diagram illustrates the stepwise selection process of osteomyelitis patients from the MIMIC-IV database, including initial diagnosis identification, radiology report linkage, deduplication, and final cohort stratification into training, validation, and test sets.

Predictor variables

Structured predictors included age, sex, comorbidities, heart rate, systolic blood pressure, temperature, oxygen saturation, white blood cell count, creatinine, hemoglobin, glucose, platelets, albumin, blood urea nitrogen, red cell distribution width, and calcium. We also derived variables such as age squared, a log-transformed days-since-first-record measure, and a prior-report indicator to capture longitudinal and repeated-encounter patterns. These features reflect disease burden, systemic inflammatory state, and physiological instability, all of which are strongly linked to prognosis in severe infection(20).

Text-based predictors were generated from radiology reports using sentence-transformer embeddings. This approach converts narrative clinical text into dense vector representations, allowing the model to capture semantic context, uncertainty, and diagnostic emphasis without relying solely on manual keyword extraction(21). In the present study, the embedding features were combined with selected structured variables to form the final input matrix for prediction.

Outcome definition

The primary endpoint was one-year mortality. Mortality status was derived from the available longitudinal clinical data within the MIMIC-IV framework. The study focused on all-cause mortality within

one year rather than cause-specific death, given the practical constraints of the available data and the aim of broad prognostic risk modelling.

Model development and evaluation

We evaluated five supervised learning models: logistic regression, random forest, LightGBM, XGBoost, and a stacking ensemble. The structured predictors were standardized, and feature selection was applied to limit noise and reduce dimensionality. The final feature matrix consisted of the selected structured variables and the text-derived embeddings. To avoid optimistic bias from repeated observations, data were split at the patient level rather than at the record level. Models were trained on the training set and assessed on an independent test set, with performance measured by the area under the receiver operating characteristic curve (AUC), accuracy, F1 score, and calibration curves. Bootstrap resampling was used to estimate 95% confidence intervals for AUC.

Results

Patient Characteristics

A total of 2,934 patients with osteomyelitis were included in this study, with a 1-year mortality rate of 7.5% (n = 220). Table 1 summarizes the baseline characteristics stratified by survival status.

Table 1
Baseline characteristics of osteomyelitis patients by 1-year mortality status

Characteristic	Overall (N = 2934) ¹	Survived (N = 2714) ¹	Died (N = 220) ¹	p-value ²
Age (years)	60.5 (15.2)	59.5 (14.9)	72.3 (13.4)	0.000
Gender				0.900
Male	1,889 (64.4%)	1,746 (64.3%)	143 (65.0%)	
Female	1,045 (35.6%)	968 (35.7%)	77 (35.0%)	
Diabetes	1,756 (59.9%)	1,631 (60.1%)	125 (56.8%)	0.378
Heart failure	836 (28.5%)	739 (27.2%)	97 (44.1%)	0.000
Renal disease	1,072 (36.5%)	963 (35.5%)	109 (49.5%)	0.000
Liver disease	176 (6.0%)	145 (5.3%)	31 (14.1%)	0.000
Cancer	159 (5.4%)	141 (5.2%)	18 (8.2%)	0.084
Metastatic cancer	57 (1.9%)	48 (1.8%)	9 (4.1%)	0.032
Heart rate, max (bpm)	102.6 (9.3)	102.5 (8.8)	104.0 (14.4)	0.128
Heart rate, min (bpm)	74.2 (6.5)	74.2 (6.2)	74.6 (9.8)	0.549
SBP, max (mmHg)	144.0 (9.6)	144.2 (9.2)	141.9 (14.1)	0.019
SBP, min (mmHg)	90.5 (7.5)	90.5 (7.4)	89.7 (8.9)	0.195
Respiratory rate, max (/min)	27.3 (4.8)	27.2 (3.1)	28.7 (13.9)	0.110
Temperature, max (°C)	98.8 (3.3)	98.9 (2.4)	97.9 (8.2)	0.067
Temperature, min (°C)	97.2 (5.4)	97.3 (4.8)	96.0 (10.1)	0.057
SpO2, min (%)	92.8 (2.1)	92.8 (2.0)	92.4 (3.1)	0.028
WBC, max (×10 ³ /μL)	10.2 (6.3)	10.1 (6.2)	12.1 (7.2)	0.000
Creatinine, max (mg/dL)	1.7 (1.7)	1.7 (1.7)	1.9 (1.6)	0.024
Hemoglobin, min (g/dL)	10.2 (1.8)	10.3 (1.8)	9.5 (1.7)	0.000
Platelets, min (×10 ³ /μL)	268.0 (110.9)	270.1 (108.8)	242.8 (131.8)	0.003
Albumin, min (g/dL)	3.0 (0.3)	3.0 (0.3)	2.9 (0.4)	0.000
BUN, max (mg/dL)	28.0 (19.8)	27.1 (18.7)	39.2 (27.6)	0.000
² Welch Two Sample t-test; Pearson's Chi-squared test				

Characteristic	Overall (N = 2934) ¹	Survived (N = 2714) ¹	Died (N = 220) ¹	p-value ²
RDW, max (%)	15.1 (2.0)	15.0 (2.0)	16.3 (2.2)	0.000
Calcium, total (mg/dL)	8.8 (0.6)	8.8 (0.6)	8.6 (0.6)	0.000
¹ Mean (SD); n (%)				
² Welch Two Sample t-test; Pearson's Chi-squared test				

Data are presented as mean (SD) or n (%). SBP, systolic blood pressure; WBC, white blood cell count; BUN, blood urea nitrogen; RDW, red cell distribution width.

Non-survivors were significantly older (72.3 vs. 59.5 years, $p < 0.001$) and had higher proportions of heart failure (44.1% vs. 27.2%, $p < 0.001$), renal disease (49.5% vs. 35.5%, $p < 0.001$), liver disease (14.1% vs. 5.3%, $p < 0.001$), and metastatic cancer (4.1% vs. 1.8%, $p = 0.032$). Diabetes and cancer were not significantly different between groups ($p > 0.05$).

Among laboratory parameters, non-survivors had significantly higher WBC (12.1 vs. $10.1 \times 10^3/\mu\text{L}$, $p < 0.001$), creatinine (1.9 vs. 1.7 mg/dL, $p = 0.024$), BUN (39.2 vs. 27.1 mg/dL, $p < 0.001$), and RDW (16.3% vs. 15.0% , $p < 0.001$), and significantly lower hemoglobin (9.5 vs. 10.3 g/dL, $p < 0.001$), platelets (242.8 vs. $270.1 \times 10^3/\mu\text{L}$, $p = 0.003$), albumin (2.9 vs. 3.0 g/dL, $p < 0.001$), and serum calcium (8.6 vs. 8.8 mg/dL, $p < 0.001$). These findings indicate that patients who died within one year had greater infection severity, worse renal function, poorer nutritional status, and higher inflammatory burden at baseline.

Model performance

Figure 2. Receiver operating characteristic (ROC) curves of the five models on the test set.

[Insert Fig. 2 here]

The curves demonstrate the discriminatory performance of Logistic Regression (AUC = 0.796, 95% CI: 0.788–0.804), Stacking Ensemble (AUC = 0.795, 95% CI: 0.787–0.803), XGBoost (AUC = 0.778, 95% CI: 0.771–0.786), Random Forest (AUC = 0.760, 95% CI: 0.751–0.768), and LightGBM (AUC = 0.761, 95% CI: 0.752–0.769). The diagonal dashed line represents the performance of a random classifier (AUC = 0.5). Different line styles and gray shades distinguish the five models for clarity.

The comparative performance of the five models is summarized in Table 2. The stacking ensemble achieved the best discrimination, with an AUC of 0.795 (95% CI 0.787–0.803), accuracy of 0.814, and F1 score of 0.414. Logistic regression also performed strongly, with an AUC of 0.796, whereas the random forest, LightGBM, and XGBoost models showed slightly lower discrimination on the held-out test set. The observed AUC values were clinically meaningful and consistent with the expectation that prediction of one-year outcomes in complex inpatient populations is challenging but not impossible.

Table 2
Model performance comparison on the test set

Model	AUC	CI_lower	CI_upper	Accuracy	F1
Logistic Regression	0.796266	0.7872	0.804128	0.676931	0.491387
Random Forest	0.760349	0.751362	0.768416	0.596213	0.44929
LightGBM	0.760526	0.751753	0.769222	0.701238	0.436202
XGBoost	0.778405	0.77057	0.786179	0.774225	0.419474
Stacking Ensemble	0.794604	0.786902	0.802893	0.814388	0.414064

AUC, area under the receiver operating characteristic curve; CI, confidence interval.

Logistic Regression achieved the highest AUC of 0.796 (95% CI: 0.787–0.804), outperforming all other models including Stacking Ensemble (AUC 0.795, 95% CI: 0.787–0.803), XGBoost (AUC 0.778, 95% CI: 0.771–0.786), Random Forest (AUC 0.760, 95% CI: 0.751–0.768), and LightGBM (AUC 0.761, 95% CI: 0.752–0.769). The superior performance of Logistic Regression suggests that the selected features—encompassing demographics, vital signs, comorbidities, laboratory parameters, RDW, and calcium—exhibit a predominantly linear relationship with 1-year mortality in this cohort. The Stacking Ensemble achieved the highest accuracy (80.8%), while Logistic Regression provided the most balanced F1 score (0.491).

Figure 3. Calibration curves of the five models on the test set.

[Insert Fig. 3 here]

The calibration curves (Fig. 3) demonstrated that Logistic Regression and Stacking Ensemble were the best-calibrated models, with predicted probabilities closely matching observed event rates across the risk spectrum. This indicates that the predicted probabilities from these models reliably reflect the actual mortality risk, supporting their potential clinical utility.

Feature Importance

Figure 4. Top 10 static and top 10 BERT dimension feature importance from the XGBoost model.

[Insert Fig. 4 here]

Static features are shown in black, and BERT dimensions are shown in gray. Among static features, age, RDW, WBC, creatinine, albumin, and calcium ranked as the most important predictors. The presence of BERT dimensions among the top features indicates that radiology report text contains complementary prognostic information beyond routine clinical variables.

Feature importance analysis from the XGBoost model (Fig. 4) identified age, RDW, WBC, creatinine, albumin, and calcium among the top predictors of 1-year mortality. Notably, RDW—a marker of systemic inflammation and nutritional status—ranked within the top five, highlighting its potential as a prognostic biomarker in osteomyelitis. Several BERT-derived text dimensions also appeared in the top 15 features, indicating that radiology report text contains complementary prognostic information beyond routine clinical variables.

Interpretation of BERT Dimensions

Figure 5. Heatmap showing the association between top 10 BERT dimensions and predefined clinical concepts.

[Insert Fig. 5 here]

Darker shades indicate stronger associations. BERT_371 showed strong association with fracture/trauma concepts; BERT_379, BERT_366, and BERT_368 were associated with vascular access procedures; and BERT_360 and BERT_368 captured pulmonary/pleural abnormalities. The presence of right and left across multiple dimensions indicates that the model attends to lesion laterality.

To enhance model interpretability, we performed back-mapping analysis of the 10 most important BERT dimensions (Fig. 5). Several clinically meaningful patterns emerged:

- **BERT_371** was heavily associated with keywords indicating fracture and trauma (fracture, distal, tibial, fibula), suggesting that the model captures the underlying traumatic etiology of osteomyelitis.
- **BERT_379, BERT_366, and BERT_368** strongly activated on reports mentioning vascular access procedures (PICC, hemodialysis, catheter), reflecting the high prevalence of indwelling lines in patients with chronic comorbidities.
- **BERT_360 and BERT_368** were associated with pulmonary abnormalities (pleural effusion, atelectasis, consolidation), consistent with the known respiratory complications in immobilized or critically ill osteomyelitis patients.
- The frequent appearance of right and left across multiple dimensions indicates that the model attends to **laterality** of the lesion—a clinically relevant attribute often overlooked in structured feature sets.

These findings demonstrate that the BERT-derived dimensions encode clinically meaningful, interpretable semantics that are complementary to the structured features, providing empirical support for the utility of radiology report text in prognostic modeling.

Discussion

Our study demonstrates that a multimodal machine learning framework can predict one-year mortality in patients with osteomyelitis with clinically meaningful accuracy by integrating radiology report text and

structured clinical variables. This finding is relevant because osteomyelitis is a heterogeneous condition in which prognosis depends on a complex interaction among infection burden, host factors, organ dysfunction, and complications. A prediction model that incorporates both quantitative clinical measurements and narrative imaging interpretation is therefore better aligned with real-world clinical reasoning than one based solely on structured variables(5).

The most important contribution of this work is the recognition that radiology reports are not merely administrative documentation. They are a rich and underused source of prognostic information. In many clinical settings, radiology reports contain specialized language about disease extent, progression, suspicion of complications, and certainty of diagnosis. These details may reflect the trajectory of illness even before they are translated into coded diagnoses or laboratory alerts. Our results support the view that narrative radiology text can be transformed into clinically meaningful features and incorporated into predictive modelling.

The study also illustrates the value of combining heterogeneous sources of information. Structured variables capture measurable physiology and comorbidity burden, whereas text embeddings capture semantic and contextual information that is often invisible to conventional modelling approaches. The combination of these two modalities allows the model to incorporate both quantitative severity and qualitative interpretation. Such integration is especially important in complex infectious diseases, where prognosis is determined not only by isolated biomarkers but also by the broader clinical context.

Several limitations should be acknowledged. First, the study was retrospective and used a single public database, which may limit generalizability to other institutions and healthcare systems. Second, the outcome was binary one-year mortality and did not capture the full complexity of patient trajectories, quality of life, or treatment response. Third, although the text embeddings performed well, future work should evaluate whether biomedical domain-specific language models can further improve performance and interpretability(6). Finally, prospective external validation in real clinical workflows will be necessary before such models can be considered suitable for decision support.

Conclusions

In summary, this study demonstrates that integrating structured clinical features with radiology report text can improve prediction of one-year mortality in patients with osteomyelitis. The multimodal framework achieved clinically meaningful discrimination and supports the use of artificial intelligence methods for risk stratification in musculoskeletal infection. These findings provide a strong foundation for future prospective validation, clinical implementation, and broader application of multimodal clinical prediction models.

Declarations

Ethics approval and consent to participate

This study utilized data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. Access to MIMIC-IV was granted after completion of the required training and credentialing (Certification ID: 77653262). The MIMIC-IV database received ethical approval from the Institutional Review Board of Beth Israel Deaconess Medical Center, and the need for individual patient informed consent was waived because the data are de-identified and publicly available. This study was conducted in accordance with the Declaration of Helsinki and followed the ethical standards for research involving human subjects. The establishment of MIMIC-IV was approved by the institutional review boards of the Beth Israel Deaconess Medical Center (Boston, MA, USA) and the Massachusetts Institute of Technology (Cambridge, MA, USA), and the database is compliant with the Health Insurance Portability and Accountability Act (HIPAA) privacy rules.

Authors' contributions

Yuhang Jin: Conceptualization, Methodology, Formal Analysis, Writing.

Zhipeng Peng: Data Curation, Software, Validation, Writing.

Xuesong Yan: Investigation, Resources, Writing.

Jiawei Li: Methodology, Validation, Writing.

Zhuohang Wu: Data Curation, Formal Analysis, Writing.

Yongxian Zhang: Supervision, Project Administration, Writing.

All authors read and approved the final manuscript.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the Shandong Provincial Natural Science Foundation, China (Grant No. ZR2025LMB024). The funder had no role in the study design, data collection, analysis, interpretation of data, writing of the manuscript, or the decision to submit the manuscript for publication.

References

1. Kremers HM, Nwojo ME, Ransom JE, Wood-Wentz CM, Melton LJ, Huddlestone PM. Trends in the epidemiology of osteomyelitis: a population-based study, 1969 to 2009. *J Bone Joint Surg Am*. 2015;97(10):837-45.

2. Tice AD, Hoaglund PA, Shoultz DA. Outcomes of osteomyelitis among patients treated with outpatient parenteral antimicrobial therapy. *The American Journal of Medicine*. 2003;114(9):723-8.
3. Arnold SR, Giordano V, Obremsky WT, Schwarz EM, Smeltzer MS, Veis DJ, et al. Osteomyelitis. *Nature reviews Disease primers*. 2026;12(1).
4. Lipsky BA, Berendt AR, Cornia PB, Pile JC, Peters EJ, Armstrong DG, et al. 2012 Infectious Diseases Society of America clinical practice guideline for the diagnosis and treatment of diabetic foot infections. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. 2012;54(12):e132-73.
5. Liu Y, Zheng Y, Ding S. Development and validation of a prognostic nomogram model for severe osteomyelitis patients. *Scientific reports*. 2025;15(1):318.
6. Liang L, Chen Z, He Z, Tao H, Chen Y, Liu T. Red cell distribution width (RDW) is independently associated with all-cause mortality in adult patients with osteomyelitis admitted to the intensive care unit. *PloS one*. 2025;20(9):e0332211.
7. Liu Y, Zheng Y, Ding S. Association between red blood cell distribution width and mortality in severe osteomyelitis patients. *Scientific reports*. 2025;15(1):360.
8. Liu Y, Zheng Y, Ding S. The correlation between serum calcium levels and prognosis in patients with severe acute osteomyelitis. *Frontiers in immunology*. 2024;15:1378730.
9. Huang Y, Jiang K, Liu Y, Mo W, Zeng Y, Yang Y, et al. Association between red blood cell distribution width to serum calcium ratio and all-cause mortality in critically ill patients with diabetic nephropathy: a retrospective analysis of the MIMIC-IV database. *BMC nephrology*. 2025;26(1):475.
10. Yasin P, Dong S, Aizezi Z, Yimit Y, Yusufu A, Yakufu M, et al. Explainable machine learning for differential diagnosis of diabetic foot infection and osteomyelitis: a two-center study and clinically applicable web calculator using routine blood biomarkers. *BMC medical informatics and decision making*. 2025;25(1):420.
11. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-30.
12. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *Jama*. 2018;319(13):1317-8.
13. Jung SJ, Seo JW. A transformer-based embedding approach to developing short-form psychological measures. *Frontiers in psychology*. 2025;16:1640864.
14. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE journal of biomedical and health informatics*. 2018;22(5):1589-604.
15. Boag W, Doss D, Naumann T, Szolovits P. 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*. 2018;2017:26-34.
16. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. *International journal of medical informatics*. 2015;84(12):1039-47.

17. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of biomedical informatics*. 2015;58:156-65.
18. Boquet-Pujadas A, Anagnostakis F, Yang Z, Tian YE, Duggan MR, Erus G, et al. Multi-organ AI Endophenotypes Chart the Heterogeneity of Pan-disease in the Brain, Eye, and Heart. *medRxiv : the preprint server for health sciences*. 2025.
19. Lipton Z, Kale D, Elkan C, Wetzel R. Learning to Diagnose with LSTM Recurrent Neural Networks. 2015.
20. Bayly H, Tripodis Y, Lenio S, Patil P, for the Alzheimer's Disease Neuroimaging I. Reducing overconfident errors in clinical prediction models. *BMC medical informatics and decision making*. 2026.
21. Cammarasana S, Patanè G. Graph representation and learning for drug administration prediction. *BMC medical informatics and decision making*. 2026.

Figures

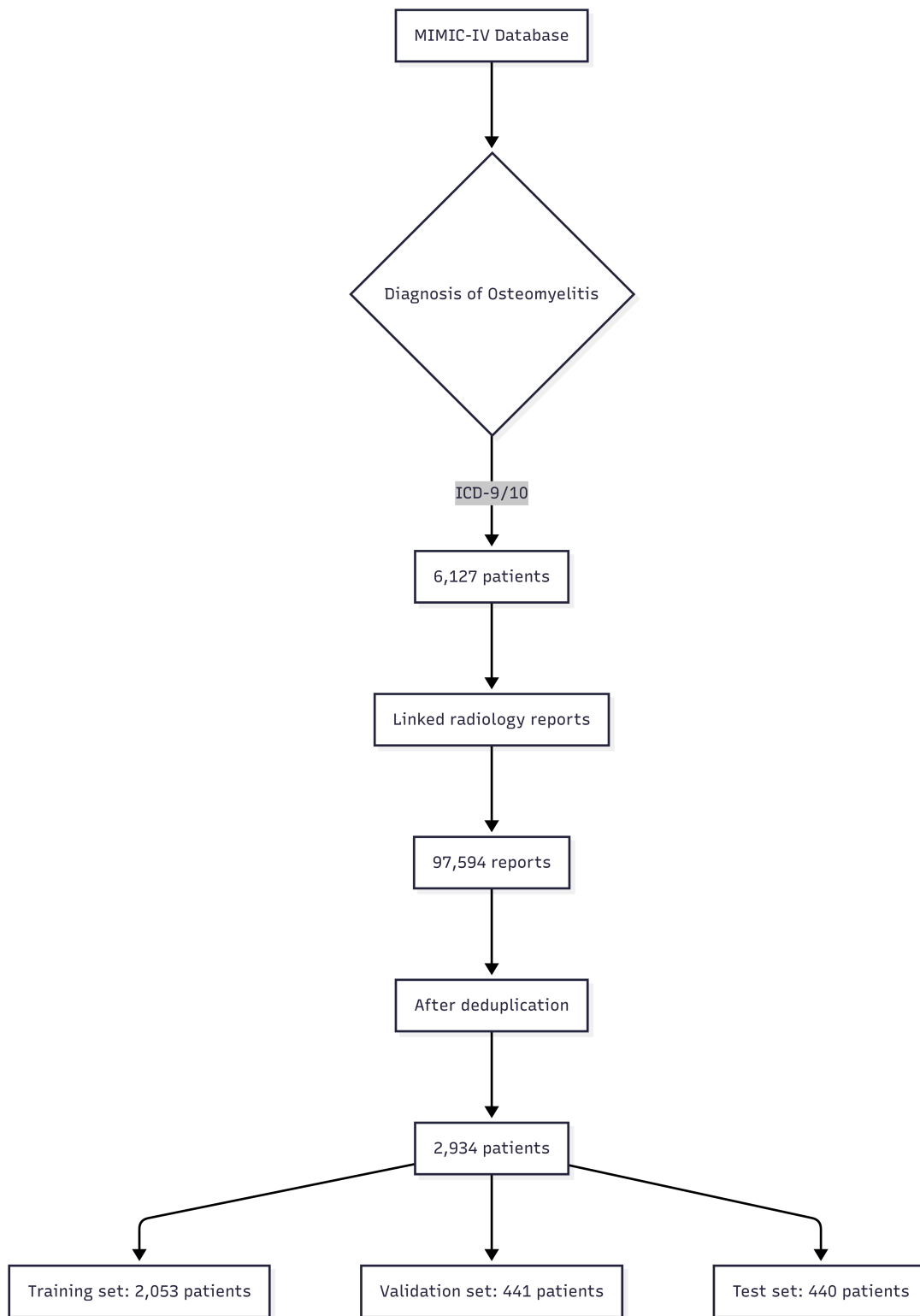


Figure 1

Patient flow diagram.

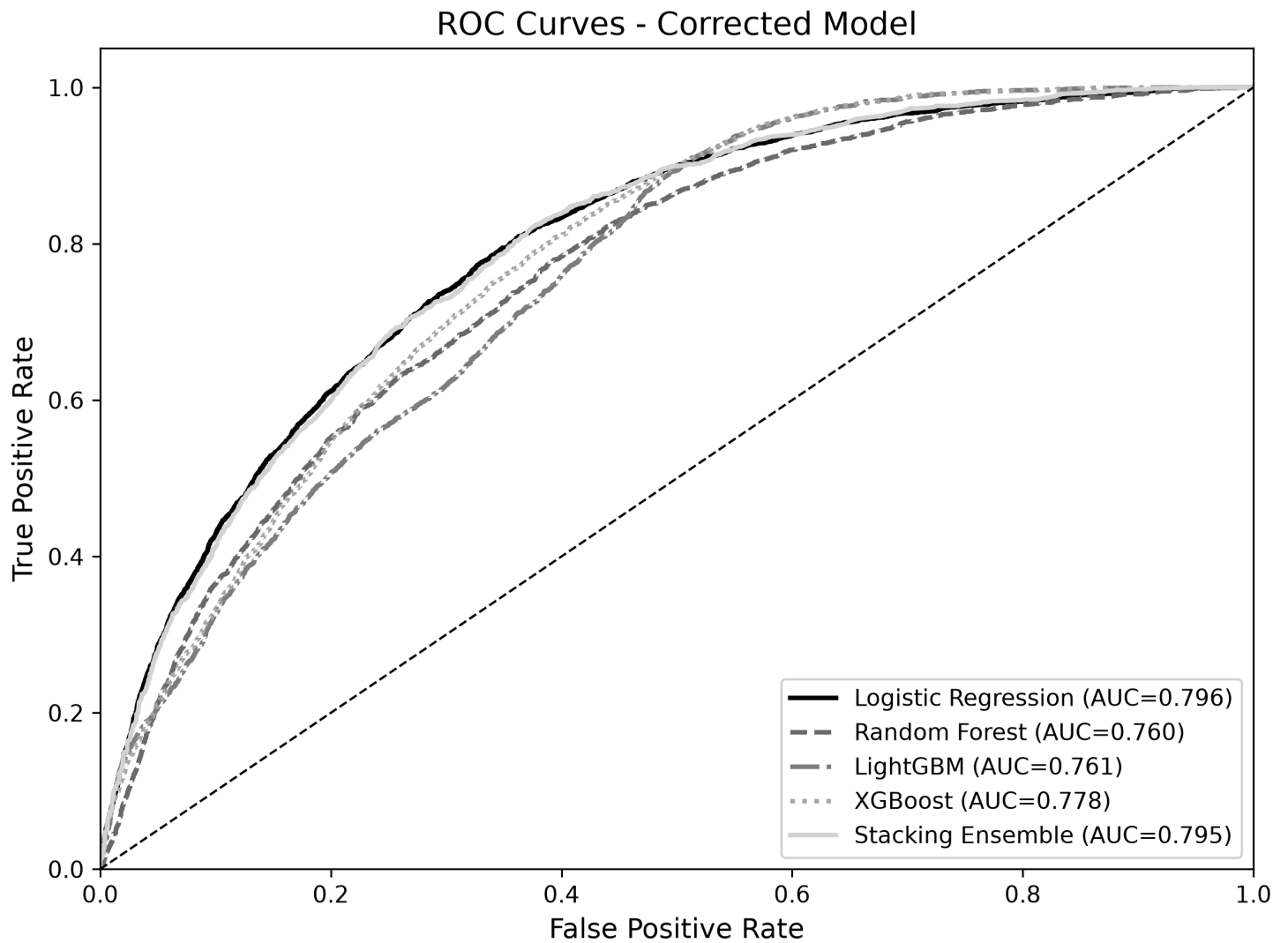


Figure 2

Receiver operating characteristic (ROC) curves of the five models on the test set.

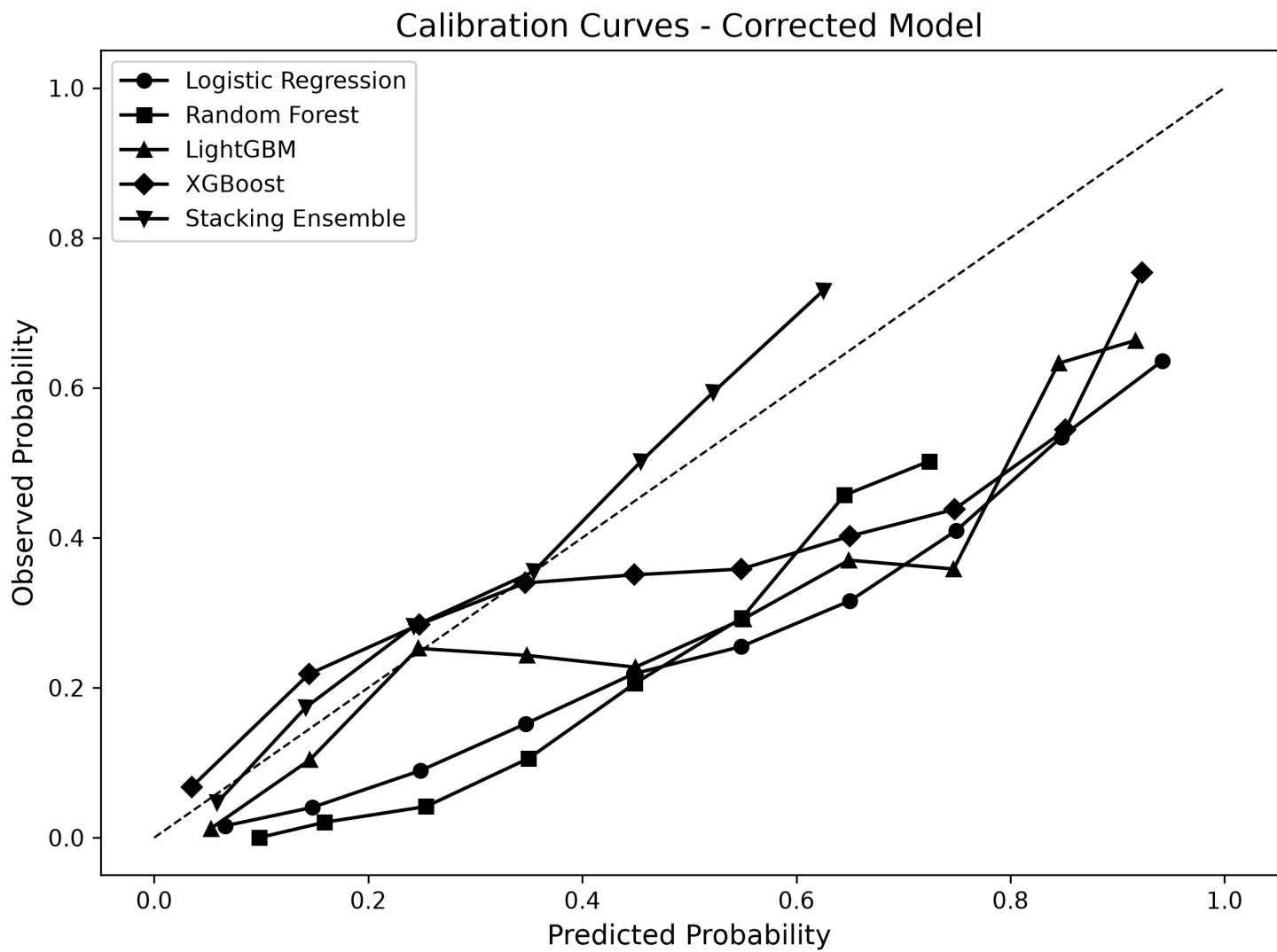


Figure 3

Calibration curves of the five models on the test set.

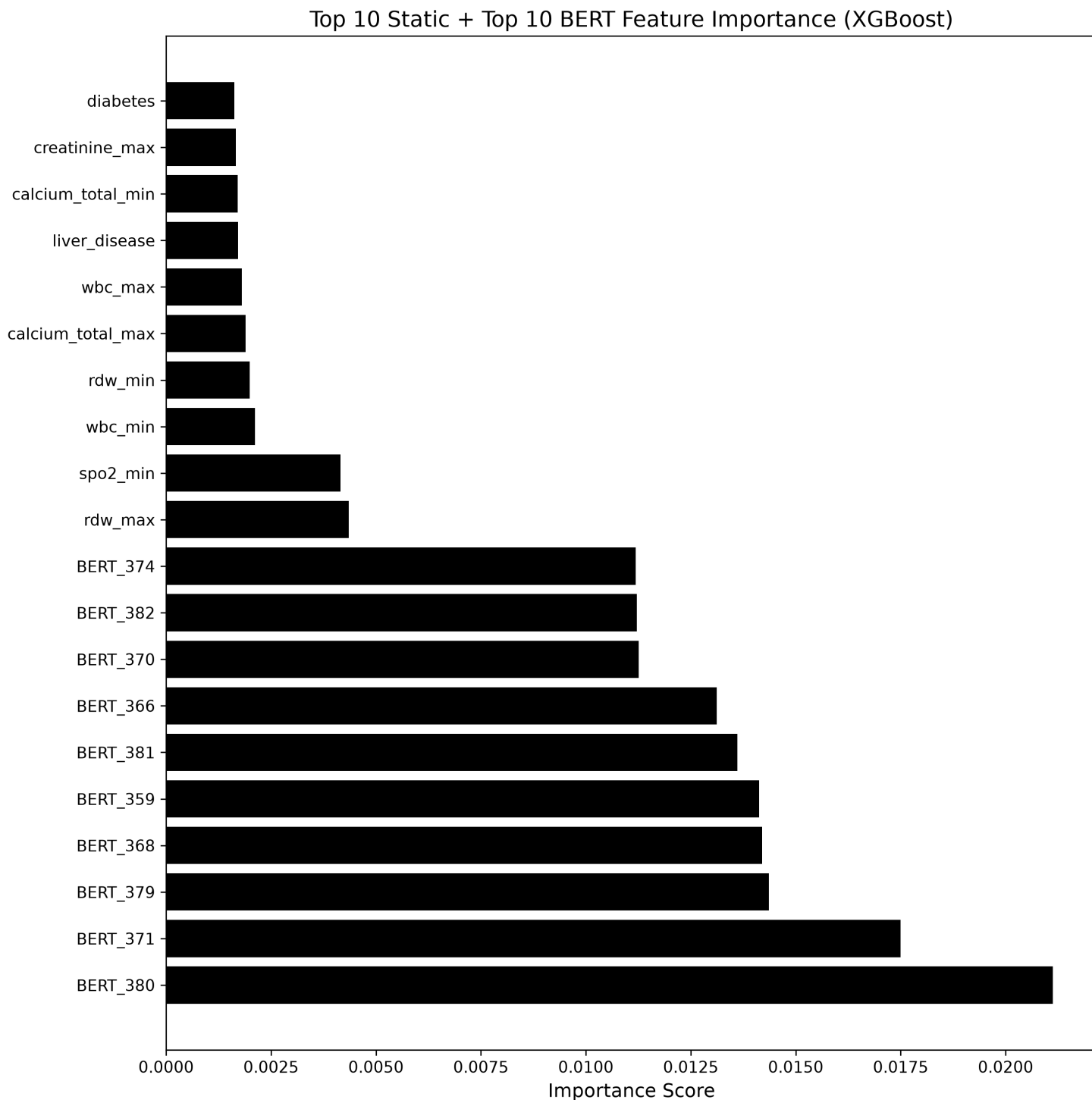


Figure 4

Top 10 static and top 10 BERT dimension feature importance from the XGBoost model.

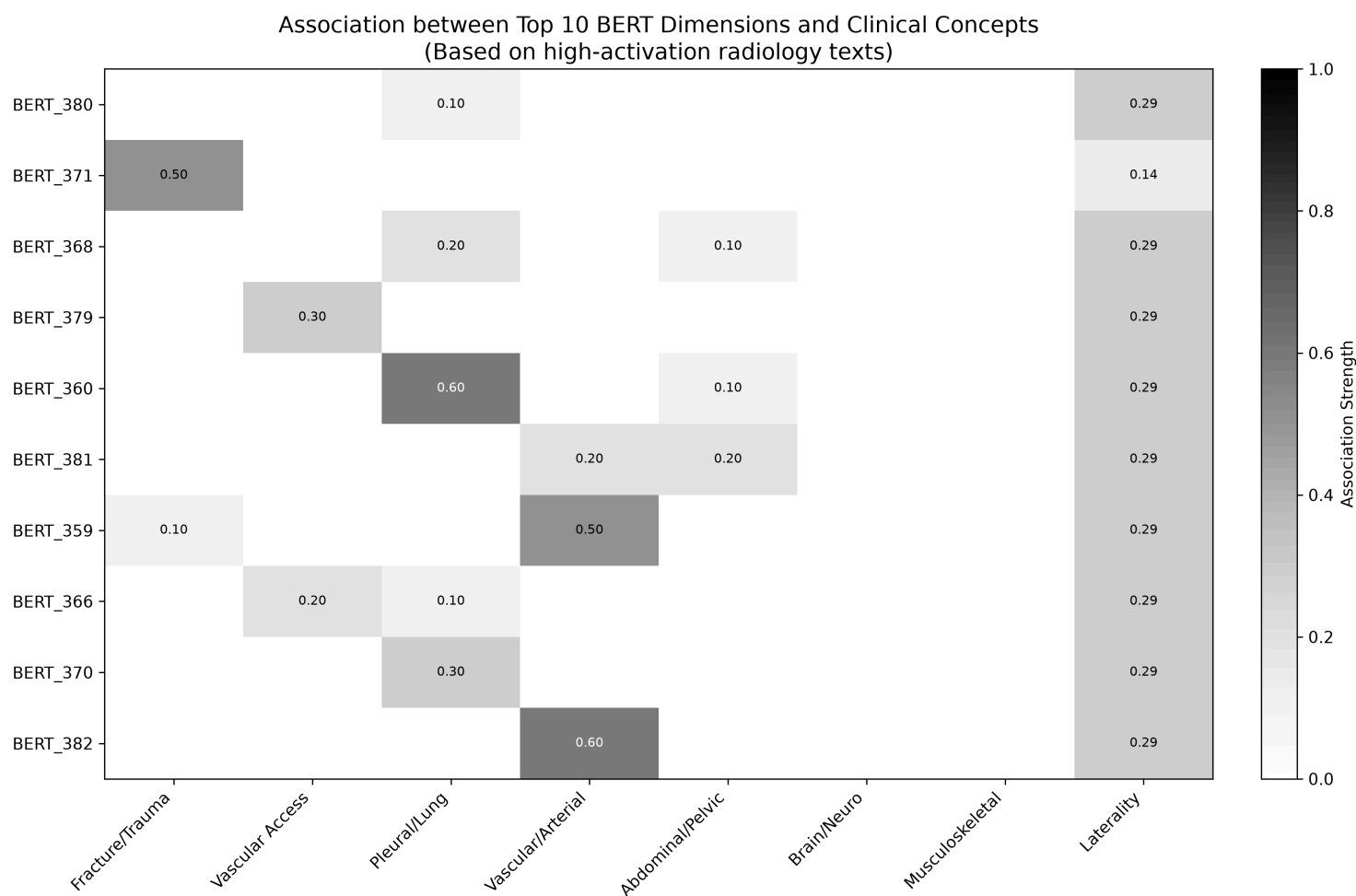


Figure 5

Heatmap showing the association between top 10 BERT dimensions and predefined clinical concepts.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [modelperformancecorrected.csv](#)
- [table1corrected.html](#)