Daniel Coblentz
VFP Summer 2025
Research Report Paper

Deep phenotyping of obstructive sleep apnea via multimodal representation learning

Daniel Coblentz[1], Itunu Ayo-Durojaiye[1], Aijuan Dong[1], Rafael Zamora-Resendiz[2], Silvia Crivelli[2]
[1]Hood College, [2]Lawrence Berkeley National Laboratory

I. ABSTRACT

Obstructive sleep apnea (OSA) is a sleep disorder characterized by intermittent pauses in breathing during sleep, and it is associated with increased cardiovascular and metabolic risks. The challenge in effective patient treatment is identifying those most at risk for related comorbidities, making improved OSA profiling and risk assessment crucial for precision healthcare. Our project utilized multimodal representation learning for deep phenotyping of sleep apnea patients. In this study, we leveraged two data modalities from electronic health records: structured clinical features and unstructured discharge summaries, to explore patient representations in a shared latent space. We integrated TabNet and ClinicalBERT encoders with late fusion, followed by Uniform Manifold Approximation and Projection and unsupervised clustering techniques. We then deep phenotyped OSA patients through unsupervised clustering, investigated factors driving differentiation into different patient subgroups, and evaluated the clinical utility of discovered representations. This project contributes new knowledge to the scientific community regarding multimodal representation learning for healthcare data and offers improved strategies for comorbidity risk stratification.

II. INTRODUCTION

Obstructive sleep apnea (OSA) is a common sleep disorder characterized by repeated episodes of reduced (hypopnea) or complete (apnea) blockage of the upper airway during sleep. This results in intermittent oxygen desaturation, frequent sleep disruptions, and subsequent daytime impairments, including excessive fatigue and cognitive deficits. In the United States, OSA affects approximately 25–30% of adult males and 9–17% of adult females [1], with global estimates indicating nearly one billion affected individuals [2]. While studies have highlighted an association between OSA and increased risk of cardiovascular disease [3], current diagnostic practices often fail to identify which patients are most vulnerable to these comorbid conditions. Polysomnography (PSG), the current standard for diagnosing sleep disorders including OSA, provides detailed physiological data but is expensive, time-consuming, and geographically inaccessible for many patients. An alternative test, home sleep apnea testing (HSAT), offers improved accessibility but with decreased accuracy, particularly in identifying mild or complex presentations of sleep disorders.

Over the past two decades, the widespread adoption of electronic health records (EHRs) has curated large datasets that combine clinical data (demographics, medication codes, lab results) with patient notes (discharge summaries). While originally designed for administrative use, EHRs have emerged as a valuable resource for machine learning applications aimed at predicting patient trajectories and treatment outcomes. One especially promising application is deep phenotyping [4], which uses machine learning to find meaningful patient groups based on shared traits in complex data, such as EHRs.

To perform deep phenotyping effectively, researchers often rely on unsupervised learning methods, such as clustering algorithms to group similar data points together without predefined labels to uncover hidden structure in the data. However, traditional approaches like K-means [5] and Gaussian Mixture Models [6] struggle with the high dimensionality, sparsity, and heterogeneity of real-world EHRs. These limitations can lead to unstable or overlapping clusters that fail to reflect clinically actionable subgroup identification.

Given these limitations of traditional clustering approaches, researchers have increasingly turned to advanced machine learning techniques that can better handle the complexity of clinical data. Recent advances in multimodal representation learning [7, 8] offer a promising alternative for handling this heterogeneous data. It aims to help machines understand complex interactions by combining information from different sources, like images, text, and audio. By using the strengths of each modality, multimodal learning allows AI systems to build stronger and richer internal representations. This alignment supports downstream tasks like patient clustering and subgroup identification, ultimately enhancing the effectiveness of clinical phenotyping.

For example, Khadanga et al [9] demonstrated that combining structured features with unstructured clinical text using a deep learning framework significantly improved predictive performance in clinical tasks such as mortality prediction and 30 day readmission, demonstrating the value of multimodal fusion in clinical practice. Their approach illustrates the potential of multimodal learning to capture a broader clinical context and drive improved outcomes in patient stratification.

Inspired by this direction, we adopt a dual-modality framework that embeds structured EHR features and unstructured discharge summaries into a shared latent space representation. Our approach uses late fusion [10] due to its empirical simplicity and stability. By bringing these data modalities together, our goal is to identify distinct OSA-related subgroups and uncover hidden patterns that may support personalized care and comorbidity risk stratification.

III. MATERIALS & METHODS

A. Data set

The Medical Information Mart for Intensive Care IV (MIMIC-IV) database was selected as the primary data source for this project [11]. It contains de-identified electronic health records (EHRs) from patients admitted to the emergency department or intensive care units of Beth Israel Deaconess Medical Center between 2008 and 2019. The most recent release, MIMIC-IV v2.2 (January 2023), includes data on 299,712 patients across 431,231 admissions. The dataset, excluding imaging data such as radiology and X-rays, occupies 75 GB of storage.

B. Computing resource platform

The computational demands of this project for training deep learning models on multimodal EHR data require substantial processing power. Tasks such as learning joint representations across large patient datasets are highly resource-intensive. The National Energy Research Scientific Computing Center (NERSC), a high-performance computing facility at Lawrence Berkeley National Laboratory, houses Perlmutter, a supercomputer with 3,072 CPU-only nodes and 1,792 GPU-accelerated nodes. This infrastructure provides the scalability necessary for processing large clinical datasets, efficiently training representation models, and performing high-throughput clustering and dimensionality reduction.

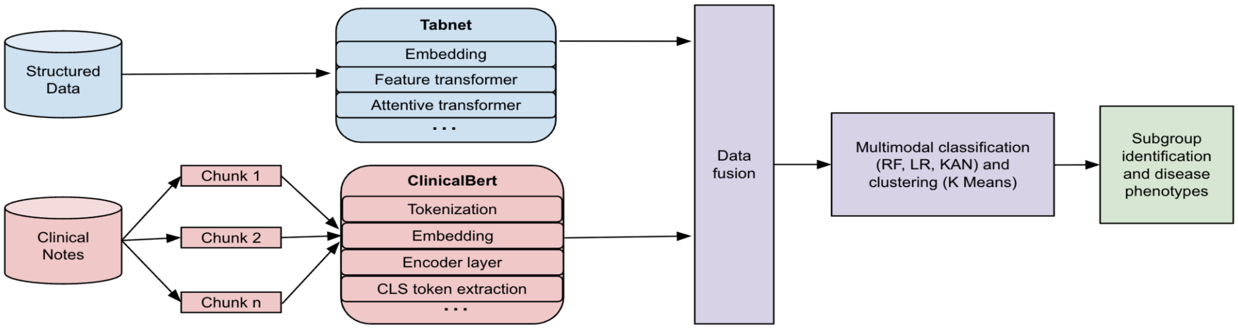

FIG. 1. Structured data and clinical notes are encoded separately using TabNet and ClinicalBERT. Their embeddings are fused for multimodal classification and clustering to identify patient subgroups and disease phenotypes.

C. Research approach

Our research approach consists of five main phases (Fig.1):

1. Feature Extraction and Preprocessing

Clinical features were extracted from two data modalities within MIMIC-IV: structured EHR data (e.g., demographics, medications, lab results) and unstructured discharge summaries. To ensure compatibility across modalities, structured features were standardized using min-max normalization, while discharge summaries were segmented into smaller chunks to respect token limits during transformer-based encoding. Patient records were filtered to include only adult patients with documented diagnoses of obstructive sleep apnea (ICD-9: 327.23, ICD-10: G47.33), heart failure (ICD-9: 428.x, ICD-10: I50.x), or both conditions (combined cohort), with at least two admissions to ensure diagnostic consistency. This filtering resulted in a final cohort of 33,977 patients and a feature set of 1,344 variables, yielding a diverse representation of cardiovascular and respiratory phenotypes.

2. Unimodal Representation Learning

For structured data, a TabNet model [12], a deep learning architecture for tabular data, was trained to generate 128-dimensional embeddings. For unstructured text, ClinicalBERT [13], a transformer model pretrained on medical notes, generated 768-dimensional embeddings from discharge summaries, which were then passed through a TabNet model to reduce them to 128 dimensions, ensuring consistency with the structured representation.

3. Multimodal Fusion

To integrate the two modalities, we concatenated the structured and unstructured embeddings into a shared feature space, forming fused patient representations. This late fusion approach enables the model to retain modality-specific strengths while allowing downstream analysis to benefit from complementary information.

4. Dimensionality Reduction, Clustering, and Classification

Uniform Manifold Approximation and Projection UMAP [14] was applied for dimensionality reduction and visualization. Unsupervised clustering was performed using KMeans and evaluated with Silhouette Score, Adjusted Rand Index (ARI), and Davies–Bouldin Index (DBI). Classification models (Random Forest (RF), Logistic Regression (LR), Kolmogorov–Arnold Network (KAN) [15]) were trained on structured, unstructured, and fused modalities and evaluated using Area Under the Receiver Operating Characteristic Curve (AUROC).

5. Subgroup Interpretation and Evaluation

Lastly, we conducted latent space analyses: 1) Deep phenotyping through unsupervised clustering to identify patient subgroups, 2) investigation of factors driving differentiation between patient groups, and 3) clinical validation of discovered phenotypes through comparison with known OSA risk factors.

IV. RESULTS

A. Clustering performance

We evaluated the learned embeddings across three configurations: structured-only, unstructured-only, and a fused representation. UMAP was applied to each modality prior to assess clustering quality in the latent space.

Modality	dimensions	ARI	Silhouette Score	DBI
Structured	128	0.25	0.49	0.66
Unstructured	128	0.19	0.48	0.68
Multimodal	256	0.09	0.74	0.48

Table I. Clustering performance comparison across data modalities. Structured, unstructured and fused embeddings (ARI: -1 to 1, Silhouette: -1 to 1, DBI: 0 to ∞; higher ARI and Silhouette values indicate better clustering, lower DBI is favorable).

Table I summarizes the clustering performance across different data modalities. The fused multimodal representations demonstrated superior clustering quality compared to individual modalities, achieving the highest Silhouette Score (0.74) and lowest Davies-Bouldin Index (0.48), indicating stronger internal consistency and better separation between patient groups. While structured data alone produced moderate clustering performance, unstructured clinical notes showed weaker separation, likely reflecting the inherent noise in free-text data. These results confirm that combining structured and unstructured data captures complementary patient information, resulting in more robust representations.

B. Classification performance

To validate the clinical utility of our discovered patient representations, we evaluated their performance in downstream predictive tasks. This assessment serves two purposes: first, to ensure that multimodal fusion preserves predictive power compared to single modalities, and second, to identify which modality most strongly drives the observed clustering patterns.

Model	Structured		Unstructured		Multimodal	
	ROC AUC	Runtime	ROC AUC	Runtime	ROC AUC	Runtime
RF	0.878	00:35	0.713	00:58	0.884	01:07
LR	0.890	00:01	0.721	00:03	0.886	00:04
KAN	0.890	02:00	0.736	04:27	0.756	05:00

Table II: ROC AUC and runtime comparison across models and modalities.

Structured features consistently provided strong baseline performance across all classifiers, with LR and KAN achieving 0.890 AUROC. The multimodal fusion maintained comparable performance while incorporating richer contextual information, with especially strong performance from LR (0.886 AUROC). Notably, unstructured data alone showed lower predictive power, reinforcing the value of multimodal integration for clinical applications.

C. Subgroup analysis

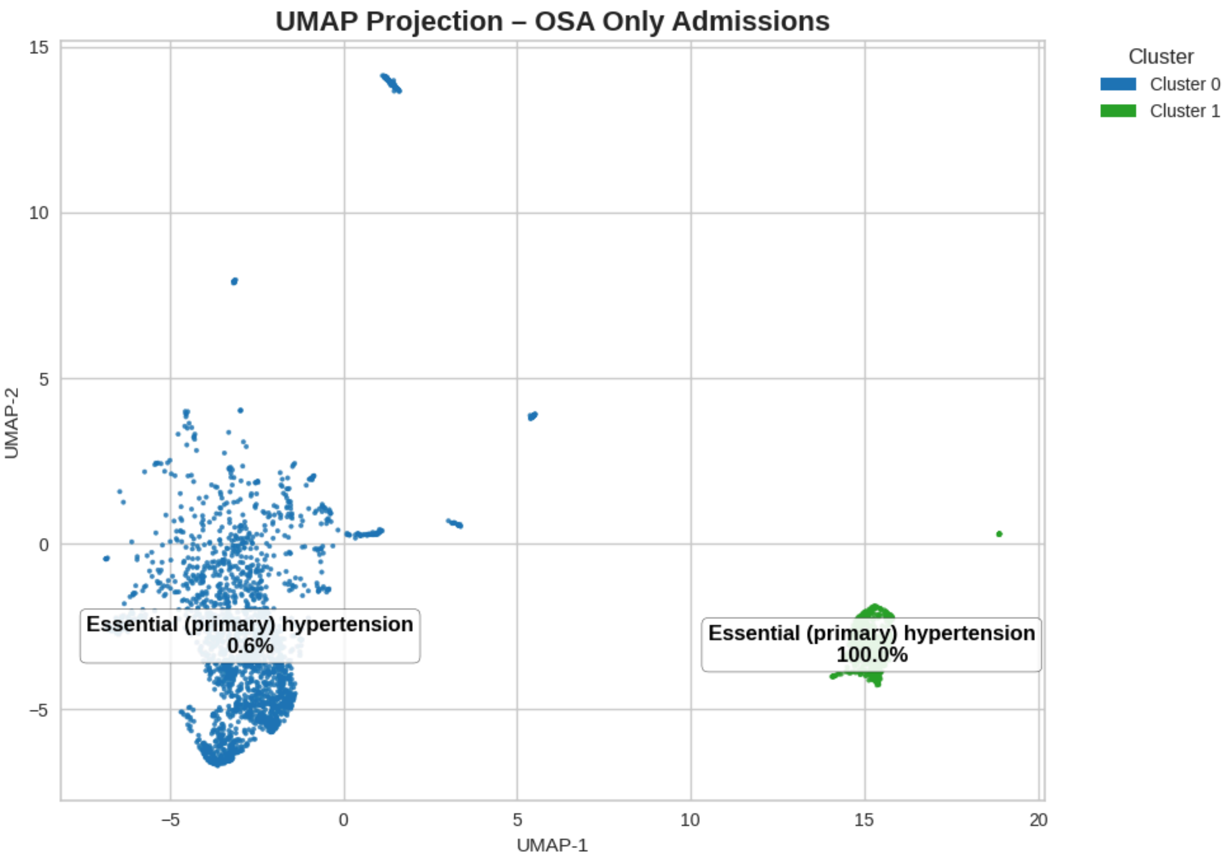

FIG. 2. Fused multimodal embeddings of OSA-only patients clustered into two subgroups (Cluster 0 and Cluster 1). Each cluster is annotated with the most discriminative feature, Essential (primary) hypertension, along with the percentage of patients in that cluster exhibiting the condition.

To characterize subgroup-specific clinical traits, we analyzed feature importance across the fused multimodal representations. The clustering algorithm identified two distinct OSA subgroups: Cluster 0 (admissions = 6,191; 90.1%) and Cluster 1 (admissions = 681; 9.9%). Analysis of the fused embeddings revealed that structured clinical features dominated the top rankings, with the first 45 most important features all originating from the structured modality. However, the fused representations could not be directly decoded to identify specific clinical features.

To identify the key discriminative clinical features, we conducted a systematic analysis of prevalence differences across all 1,340 structured features between the two clusters. Essential hypertension emerged as the most discriminative feature, with Cluster 1 exhibiting 100% prevalence versus 0.6% in Cluster 0 (difference = 99.4%). This finding was corroborated by TabNet's interpretable feature importance analysis, which independently identified hypertension-related features as top contributors to the structured representation.

Given its elevated hypertension prevalence, higher BMI, and other overweight-associated characteristics, Cluster 1 was initially suspected to have worse clinical outcomes. However, survival analysis revealed the opposite: the larger Cluster 0 despite its lower cardiovascular comorbidity burden demonstrated poorer long-term survival. This unexpected finding highlights the value of unsupervised clustering in revealing outcome patterns not apparent from baseline comorbidity profiles alone.

D. Survival analysis

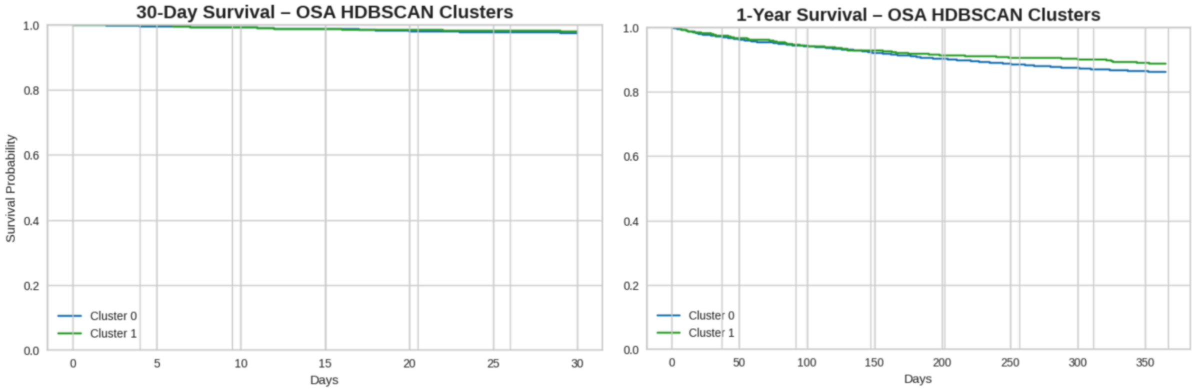

FIG. 3. Survival curves at 30-day (left) and 365-day (right) time intervals, modeled for the same two OSA subgroups shown in Fig. 2.

We evaluated differences in clinical outcomes across the two unsupervised patient subgroups using Kaplan–Meier survival analysis, stratified by HDBSCAN-derived cluster membership and estimated via the Kaplan–Meier estimator [16]. Survival probabilities were assessed at 30, 60, 180 days, and 1 year post-admission, with patients censored at the last known follow-up if no mortality event had occurred.

The log-rank test indicated a statistically significant difference in long-term survival ($p < 0.05$), with Cluster 0 exhibiting lower survival probabilities over the 1-year period compared to Cluster 1. In the short term (30 days), survival probabilities were similar between clusters, but divergence between survival lines began after 180 days. These findings underscore that the cluster with a lower hypertension prevalence and larger population size was, unexpectedly, at higher risk of long-term mortality.

V. DISCUSSION & FUTURE WORK

Our results demonstrate that multimodal fusion can effectively identify clinically meaningful OSA subgroups that would be missed by traditional single-modality approaches. We identified a distinct subgroup representing 22.3% of patients, characterized by near-universal hypertension compared to less than 1% in the other group highlighting substantial clinical heterogeneity within the OSA population. Surprisingly, the larger subgroup, despite having a much lower prevalence of hypertension, exhibited a higher long-term mortality rate. These findings emphasize the need for more tailored risk-stratification approaches.

The improved clustering performance when combining structured and unstructured data (Silhouette Score: 0.74 vs ~0.49) suggests these modalities capture different, but complementary, aspects of patient presentations, which aligns with how clinicians naturally integrate quantitative data with clinical narratives.

In addition to the late fusion approach we conducted an exploration of contrastive learning [17] that didn't outperform simple late fusion; this likely reflects the complexity of optimizing joint embedding spaces rather than fundamental limitations of the approach. With refined architectures, larger batch sizes, and more sophisticated negative sampling strategies, contrastive learning remains a promising direction for future work.

Our study has important limitations, including its single-center design and focus on ICU patients, which may not represent the broader OSA population. Future research should validate these phenotypes across multiple healthcare systems, refine contrastive learning approaches for clinical data, and investigate whether incorporating additional data sources could reveal even more clinically relevant subgroups.

VI. CONCLUSIONS

Our research demonstrates the effectiveness of multimodal representation learning for uncovering clinically meaningful patient subgroups in individuals with obstructive sleep apnea. By integrating structured and unstructured EHR data into a shared latent space, we identified distinct patient clusters and highlighted key differentiating features associated with elevated comorbidity risk. This framework supports deeper phenotyping and provides a foundation for more targeted intervention strategies. As the field continues to grow, multi-modal learning is expected to improve many areas: computer vision, natural language processing and speech recognition. These findings offer a scalable path for healthcare systems to incorporate multimodal analytics into clinical decision support and risk stratification workflows.

VII. ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Visiting Faculty Program (VFP). This work was also supported in part by the Hood College Volpe Scholars Program. I would like to gratefully acknowledge the computational resources provided by the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility at Lawrence Berkeley National Laboratory. Additionally I would like to give a special thanks to my mentors Dr. Aijuan Dong and Dr. Silvia Crivelli for their support, guidance and expertise throughout this research project.

VIII. REFERENCES

[1] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of Sleep-Disordered Breathing in adults," Am. J. Epidemiol. 177, 1006–1014 (2013).

[2] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. M. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," Lancet Respir. Med. 7, 687–698 (2019).

[3] F. J. Nieto, T. B. Young, B. K. Lind, E. Shahar, J. M. Samet, S. Redline, R. B. D'Agostino, A. B. Newman, M. M. Lebowitz, and T. G. Pickering, "Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study," JAMA 283, 1829–1836 (2000).

[4] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep representation learning of electronic health records to unlock patient stratification at scale," npj Digit. Med. 3, 96 (2020).

[5] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Inf. Sci. 622, 178–210 (2023).

[6] L. Ruan, M. Yuan, and H. Zou, "Regularized parameter estimation in high-dimensional Gaussian mixture models," Neural Comput. 23, 1605–1622 (2011).

[7] T. Sysko-Romańczuk and D. Rak, "Effective Techniques for Multimodal Data Fusion: A Comparative Analysis," Sensors 23, 2381 (2023).

[8] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in multimodal medical imaging: A review," Artif. Intell. Rev. 57, 99 (2024).

[9] S. Khadanga, I. Deznabi, H. Yang, A. Finn, and M. Syed, "Combining structured and unstructured data for predictive models: a deep learning approach," BMC Med. Inform. Decis. Mak. 20, 295 (2020).

[10] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell. 41, 423–443 (2019).

[11] A. E. Johnson, T. J. Pollard, S. Horng, L. Shen, H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," Sci. Data 9, 317 (2022).

[12] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," Proc. AAAI Conf. Artif. Intell. 35, 6679–6687 (2021).

[13] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," arXiv:1904.05342 (2019).

[14] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," J. Open Source Softw. 3, 861 (2018).

[15] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "KAN: Kolmogorov–Arnold Networks," arXiv:2404.19756 (2024).

[16] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481.

[17] S. Ketabi and D. Ramachandram, "Bridging Electronic Health Records and Clinical Texts: Contrastive Learning for Enhanced Clinical Tasks," arXiv:2505.17643 (2025).