International Journal of Medical Science and Public Health Research
ISSN : 2767-3774
SJIF: 2021: 5.456, 2022: 5.681, 2023: 6.591
IJMSPHR

TYPE Original Research
PAGE NO. 18-28
DOI 10.37547/ijmsphr/Volume07Issue02-03

Check for updates

Population-Level Oral Disease Surveillance Using Large Language Models on Clinical and Public Health Data

OPEN ACCESS

SUBMITTED 17 December 2025
ACCEPTED 19 January 2026
PUBLISHED 05 February 2026
VOLUME Vol.07 Issue02 2026

CITATION
Phan, H. T. N. ., Nguyen, T. Q. ., & Nguyen, U. . (2026). Population-Level Oral Disease Surveillance Using Large Language Models on Clinical and Public Health Data. International Journal of Medical Science and Public Health Research, 7(02), 18–28. https://doi.org/10.37547/ijmsphr/Volume07Issue02-03

COPYRIGHT
© 2026 Original content from this work may be used under the terms of the creative common's attributes 4.0 License.

Han Thi Ngoc Phan
Dentist, Pham Hung Dental Center MTV Company Limited, Pham Hung Street, Binh Chanh district, Ho Chi Minh city, Vietnam

Trang Quynh Nguyen
Dentist, Pham Hung Dental Clinic, Ho Chi Minh City, Vietnam

Uyen Nguyen
VIVA Group, 50 Tran Khac Chan, Tan Dinh Ward, District 1, Ho Chi Minh city, Vietnam

Abstract

Population-level oral disease surveillance is critical for guiding public health interventions, yet traditional systems relying solely on structured data often fail to capture contextual, behavioral, and access-to-care determinants embedded in unstructured clinical narratives. In this study, we developed a hybrid large language model (LLM) framework that integrates structured epidemiological features with embeddings derived from examination notes and survey text to improve the detection and monitoring of dental caries, periodontal disease, and tooth loss. Using the publicly available NHANES Oral Health Dataset, we compared the performance of traditional machine learning models, text-only LLM models, and our proposed hybrid approach. The hybrid model consistently outperformed all baselines, achieving higher accuracy, precision, recall, F1-score, and calibration, while maintaining equitable performance across demographic and socioeconomic subgroups. Explainability analyses revealed that combining structured and unstructured features captured clinically meaningful patterns, including

behavioral risk factors and care access barriers. Our findings suggest that hybrid LLM-based surveillance can enhance real-time population-level monitoring, identify high-risk communities, and inform preventive strategies within the U.S. public healthcare system, offering a scalable, interpretable, and equitable approach to oral health monitoring.

Keywords: Oral disease surveillance, large language models, hybrid modeling, population health, dental caries, periodontal disease, NHANES, public health informatics

Introduction

Oral diseases represent a major public health concern in the United States and globally, affecting over 3.5 billion individuals worldwide and significantly impacting quality of life, productivity, and healthcare costs. Dental caries, periodontal disease, and tooth loss are among the most prevalent oral conditions and have well-established associations with systemic diseases such as cardiovascular disease, diabetes, and adverse pregnancy outcomes. Despite their widespread prevalence, oral diseases are often underdiagnosed and underreported in public health surveillance systems due to limitations in clinical data coverage, inconsistent reporting standards, and the reliance on episodic survey collection methods. Consequently, there exists a critical need for innovative approaches to monitor oral health at the population level and to identify high-risk communities in near–real-time.

Traditional oral health surveillance methods have primarily relied on structured epidemiological datasets, including clinical examination records and nationally representative surveys such as the National Health and Nutrition Examination Survey (NHANES). While these datasets provide valuable insights into population-level disease prevalence and risk factors, they are limited in their ability to capture nuanced behavioral, environmental, and access-to-care determinants that are often embedded in unstructured clinical notes, survey responses, and community health reports. Emerging computational approaches, including machine learning and natural language processing, provide opportunities to leverage both structured and unstructured data for more comprehensive and dynamic surveillance.

Large language models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like text, capturing semantic patterns, and extracting meaningful information from unstructured narratives. In recent years, biomedical LLMs have been applied to clinical decision support, disease phenotyping, and predictive modeling in hospital and research settings. These models can identify latent patterns in textual data, such as symptom descriptions, behavioral risk factors, and clinician observations, that are not readily quantifiable using traditional structured features. Integrating LLMs with structured epidemiological data offers a promising avenue for enhancing population-level oral disease surveillance, enabling early detection, accurate prevalence estimation, and equitable identification of vulnerable subpopulations.

In this study, we propose a hybrid surveillance framework that combines structured oral health indicators, demographic and behavioral features, and LLM-derived embeddings from unstructured text. Our objective is to assess whether this integrative approach can improve the accuracy, calibration, and equity of oral disease surveillance relative to traditional methods. We further aim to evaluate the potential applicability of such a model within the U.S. public healthcare system, particularly for monitoring disparities and guiding targeted interventions.

Literature Review

Traditional Oral Health Surveillance

Population-level oral health surveillance has historically relied on structured data sources such as clinical examinations, epidemiological surveys, and administrative claims databases. The NHANES Oral Health Component has provided standardized measures of dental caries, periodontal disease, and tooth loss across representative U.S. populations. Such datasets have been instrumental in identifying trends in disease prevalence, social determinants of oral health, and disparities in access to dental care. However, structured surveillance approaches face several limitations. First, they often rely on infrequent cross-sectional surveys, limiting the ability to capture temporal trends or emerging disease hotspots. Second, structured datasets inadequately reflect behavioral, environmental, and psychosocial factors that influence

oral health. Finally, disparities in survey participation and data completeness can bias prevalence estimates, particularly among underserved populations.

Machine Learning in Oral Health Research

Recent studies have increasingly applied machine learning techniques to enhance predictive modeling of oral disease outcomes. Traditional approaches such as logistic regression, random forest, and gradient boosting have been used to predict dental caries, periodontal disease, and tooth loss using structured clinical and demographic data. These models have demonstrated moderate predictive performance, but their ability to generalize to diverse populations or incorporate nuanced contextual information remains limited. Deep learning architectures, including multilayer perceptrons and convolutional neural networks, have also been explored to capture nonlinear associations among clinical variables. While these methods improve predictive capacity, they are constrained by their reliance on structured inputs and often lack interpretability.

Natural Language Processing and Large Language Models in Healthcare

Natural language processing (NLP) techniques have been increasingly applied to extract information from unstructured clinical notes, survey narratives, and patient-reported outcomes. Early approaches utilized rule-based systems or bag-of-words representations to identify disease mentions or behavioral patterns. However, these methods are limited in their ability to capture context, negation, or semantic relationships in complex medical text.

Large language models, particularly transformer-based architectures, have demonstrated superior performance in understanding and generating natural language in biomedical domains. Models such as BioBERT, ClinicalBERT, and GPT-derived architectures have been successfully applied to disease phenotyping, clinical note summarization, and predictive modeling tasks. By generating contextual embeddings that encode semantic relationships, LLMs enable the integration of unstructured textual information into predictive frameworks. Recent studies in epidemiology have highlighted the potential of LLMs to identify emerging disease patterns from narrative data, including social media, electronic health records, and public health reports.

Hybrid Approaches and Population Health Applications

Integrating structured epidemiological data with LLM-derived embeddings represents a novel approach for population-level disease surveillance. Hybrid models can leverage both quantitative risk factors (e.g., age, smoking status, socioeconomic indicators) and contextual narrative data (e.g., symptoms, care access barriers) to improve prediction and monitoring of disease prevalence. Such approaches have shown promise in chronic disease surveillance, mental health monitoring, and outbreak detection, but their application to oral health remains limited.

The combination of structured and unstructured data is particularly advantageous for addressing disparities in oral health. Narrative data often contain subtle indicators of delayed care, cultural or linguistic barriers, and patient-reported outcomes that are underrepresented in structured surveys. Hybrid LLM models can therefore enhance the accuracy, fairness, and interpretability of oral disease surveillance, providing actionable insights for public health interventions and policy development.

Research Gap

Despite advances in machine learning and NLP for healthcare, there is a paucity of research integrating LLMs with structured oral health datasets for population-level surveillance. Most existing studies focus on either structured epidemiological modeling or clinical text mining, but rarely in a unified framework. This gap limits the ability of public health authorities to leverage narrative data in routine surveillance and hinders equitable identification of high-risk populations.

Our study addresses this gap by developing a hybrid LLM-based model for population-level oral disease surveillance, combining structured demographic and clinical features with contextual embeddings derived from examination notes and survey narratives. By evaluating the model's predictive performance, calibration, and subgroup equity, we aim to demonstrate its potential for real-world deployment in the U.S. public healthcare system.

Methodology

Data Collection

In this study, we leveraged publicly available, large-scale clinical and public health datasets to develop a population-level oral disease surveillance framework grounded in large language model methodologies. The primary data source was obtained from the Kaggle repository and originates from the National Health and Nutrition Examination Survey Oral Health Dataset; a nationally representative dataset widely used in epidemiological and public health surveillance research. This dataset integrates structured oral examination findings with demographic, behavioral, and socioeconomic survey information, making it particularly suitable for modeling oral disease patterns at the population level.

The dataset includes adult participants aged eighteen years and older and spans multiple survey cycles, thereby ensuring heterogeneity across age groups, gender, racial and ethnic backgrounds, income strata, and access-to-care profiles. Oral health outcomes captured in the dataset include dental caries experience, periodontal disease indicators such as probing depth and clinical attachment loss, tooth loss status, and self-reported oral health conditions. In addition to structured variables, the dataset contains semi-structured textual fields derived from examination remarks and survey responses that approximate narrative-style documentation commonly found in clinical and public health settings.

A detailed description of the dataset characteristics is provided in Table 1. The combination of structured clinical indicators and unstructured narrative data enables the application of large language models while maintaining epidemiological validity for population-level surveillance. The use of an open-source dataset ensures transparency, reproducibility, and ethical compliance, as all records were fully anonymized prior to public release.

Table 1. Description of the Open-Source Oral Health Dataset Used in This Study

Dataset Name	Source Repository	Sample Size	Data Modality	Key Variables
NHANES Oral Health Dataset	Kaggle (derived from CDC NHANES)	9,845 individuals	Structured and semi-structured	Dental caries status, periodontal probing depth, clinical attachment loss, tooth loss, demographic attributes, insurance coverage, smoking status, dietary patterns, oral hygiene behaviors, examination remarks

Data Preprocessing

We conducted comprehensive preprocessing to ensure data quality, internal consistency, and compatibility with large language model architectures. Structured variables were examined for missingness, distributional anomalies, and logical inconsistencies. Missing values in demographic and behavioral variables were addressed using multiple imputation strategies to preserve population-level distributions and minimize bias. Clinical outcome variables with excessive missingness or inconsistent measurement across survey cycles were excluded from downstream modeling.

To ensure comparability across survey years, clinical oral health indicators were standardized and harmonized. Continuous variables such as age, body mass index, periodontal probing depth, and attachment loss were normalized using z-score transformation. Categorical variables including education level, insurance status, and smoking behavior were harmonized across survey cycles and encoded using embedding-compatible representations.

Textual data underwent extensive normalization,

including lowercasing, removal of extraneous symbols, and selective filtering of non-informative terms. Domain-specific dental terminology was standardized using controlled vocabularies to reduce linguistic variability while preserving clinically meaningful negations and descriptors. These preprocessing steps ensured that narrative data retained semantic fidelity while remaining suitable for language model ingestion.

Feature Extraction

Feature extraction followed a dual-stream approach to capture both quantitative epidemiological signals and qualitative semantic information. From structured data, we extracted clinically relevant indicators reflecting disease presence, cumulative risk exposure, healthcare access, and demographic stratification. These features represent well-established determinants of oral health outcomes and provide a robust foundation for population-level modeling.

For unstructured data, we employed a pre-trained biomedical large language model fine-tuned on clinical and public health corpora to generate contextual embeddings. These embeddings captured latent semantic patterns associated with symptoms, clinician observations, behavioral risk descriptions, and disease severity. Sentence-level embeddings were aggregated into participant-level representations using attention-based pooling mechanisms, allowing the model to emphasize clinically salient narrative segments.

This integrated feature extraction strategy enabled comprehensive individual-level representations that combine measurable clinical indicators with nuanced contextual information embedded in narrative text.

Feature Engineering

We performed extensive feature engineering to enhance predictive performance, interpretability, and public health relevance. Composite socioeconomic vulnerability indices were constructed by integrating income, education, and insurance coverage variables. Behavioral risk scores were derived by combining smoking status, sugar consumption frequency, alcohol use, and oral hygiene practices, reflecting established oral health risk frameworks.

Temporal features were introduced to account for survey cycle variation and cohort effects, allowing the model to adjust for changes in public health policy, healthcare access, and population behavior over time. Interaction features between clinical indicators and social determinants were generated to capture disparities in oral disease burden across subpopulations.

From language model embeddings, we derived semantic coherence and topic intensity measures that reflect recurring oral health themes such as untreated caries, periodontal inflammation, and barriers to dental care. Feature selection was guided by epidemiological relevance, variance analysis, and multicollinearity assessment to ensure model stability and interpretability.

Model Development

We developed a hybrid population-level surveillance model by integrating structured epidemiological features with large language model–derived embeddings. The core architecture consisted of a transformer-based language model coupled with a neural classification layer designed to predict oral disease outcomes and prevalence patterns. The model was trained to identify associations between clinical findings, social determinants, and narrative descriptions indicative of oral disease burden.

Training was conducted using a supervised learning framework with stratified sampling to preserve outcome prevalence distributions and address class imbalance. Fine-tuning procedures were applied to adapt the language model to public health–specific terminology and survey-style narratives. Regularization techniques and early stopping were employed to prevent overfitting and enhance generalizability.

To support interpretability and surveillance utility, attention mechanisms were analyzed to identify influential features and narrative segments contributing to model predictions, enabling transparent population-level inference.

Model Evaluation

Model performance was evaluated using a held-out test set comprising twenty percent of the dataset, stratified by oral disease outcome. Predictive accuracy was

assessed using area under the receiver operating characteristic curve, precision, recall, and F1-score to ensure balanced disease detection performance. Calibration was evaluated using Brier scores to assess the reliability of prevalence estimates.

To validate population-level applicability, model-derived disease prevalence estimates were compared with established benchmarks reported in national oral health surveillance studies, demonstrating strong alignment. Subgroup analyses were conducted across age, gender, socioeconomic status, and racial and ethnic categories to assess robustness and equity. Model explainability was examined through attention visualization and feature attribution methods to identify dominant clinical, behavioral, and social drivers of oral disease predictions.

Overall, the evaluation confirms that integrating large language models with structured public health data provides a reliable, interpretable, and scalable framework for population-level oral disease surveillance.

Results

Population Characteristics and Data Overview

After completing data preprocessing and quality control procedures, the final analytical sample consisted of 9,845 adult participants drawn from multiple NHANES survey cycles. The cohort retained the demographic and socioeconomic diversity expected of a nationally representative U.S. population. Dental caries was observed in approximately forty-six percent of participants, while thirty-two percent exhibited clinical indicators consistent with moderate to severe periodontal disease. Partial or complete tooth loss was identified in twenty-eight percent of the study population, with prevalence increasing markedly with age and socioeconomic disadvantage.

The unstructured textual component of the dataset demonstrated substantial informational richness. Examination remarks and survey narratives averaged seventy-three tokens per participant and frequently contained references to oral pain, bleeding gums, delayed dental visits, insurance barriers, and clinician-observed inflammation. These narrative elements provided contextual signals not fully captured by structured variables alone, supporting the integration of large language models for enhanced population-level surveillance.

Comparative Model Performance Analysis

We conducted a comparative evaluation of multiple modeling approaches to assess their effectiveness for population-level oral disease surveillance. Baseline models included logistic regression, random forest, and gradient boosting machines trained exclusively on structured epidemiological features. A deep learning multilayer perceptron served as an additional baseline to assess nonlinear modeling capacity using engineered numerical features. We further evaluated a transformer-based large language model applied solely to unstructured textual data. These approaches were compared against our proposed hybrid model, which integrates structured clinical and demographic features with contextual embeddings generated by a biomedical large language model.

Across all oral disease outcomes, traditional machine learning models demonstrated reasonable predictive performance but exhibited limitations in sensitivity and calibration, particularly among socioeconomically vulnerable subgroups. Gradient boosting consistently outperformed logistic regression and random forest; however, its reliance on structured data constrained its ability to capture contextual risk signals embedded in narrative text. The language-model-only approach improved recall for disease detection but showed reduced precision and less stable calibration at the population level.

The hybrid model consistently achieved superior performance across all evaluation metrics. By jointly modeling epidemiological indicators and narrative-derived semantic features, the hybrid approach demonstrated enhanced discrimination, improved calibration, and greater robustness across demographic and socioeconomic strata.

Model Performance Results

Table 2 summarizes the predictive performance of each model across key oral disease outcomes. Performance is reported using area under the receiver operating characteristic curve, precision, recall, F1-score, and Brier score for calibration assessment.

Table 2. Comparative Model Performance for Oral Disease Prediction

Model	AUC	Precision	Recall	F1-score	Brier Score
Logistic Regression (Structured Only)	0.74	0.69	0.65	0.67	0.182
Random Forest (Structured Only)	0.77	0.71	0.68	0.69	0.169
Gradient Boosting (Structured Only)	0.81	0.75	0.72	0.73	0.154
Multilayer Perceptron (Structured Only)	0.79	0.73	0.70	0.71	0.161
Transformer LLM (Text Only)	0.78	0.70	0.76	0.73	0.172
Proposed Hybrid LLM Model	**0.89**	**0.83**	**0.81**	**0.82**	**0.118**

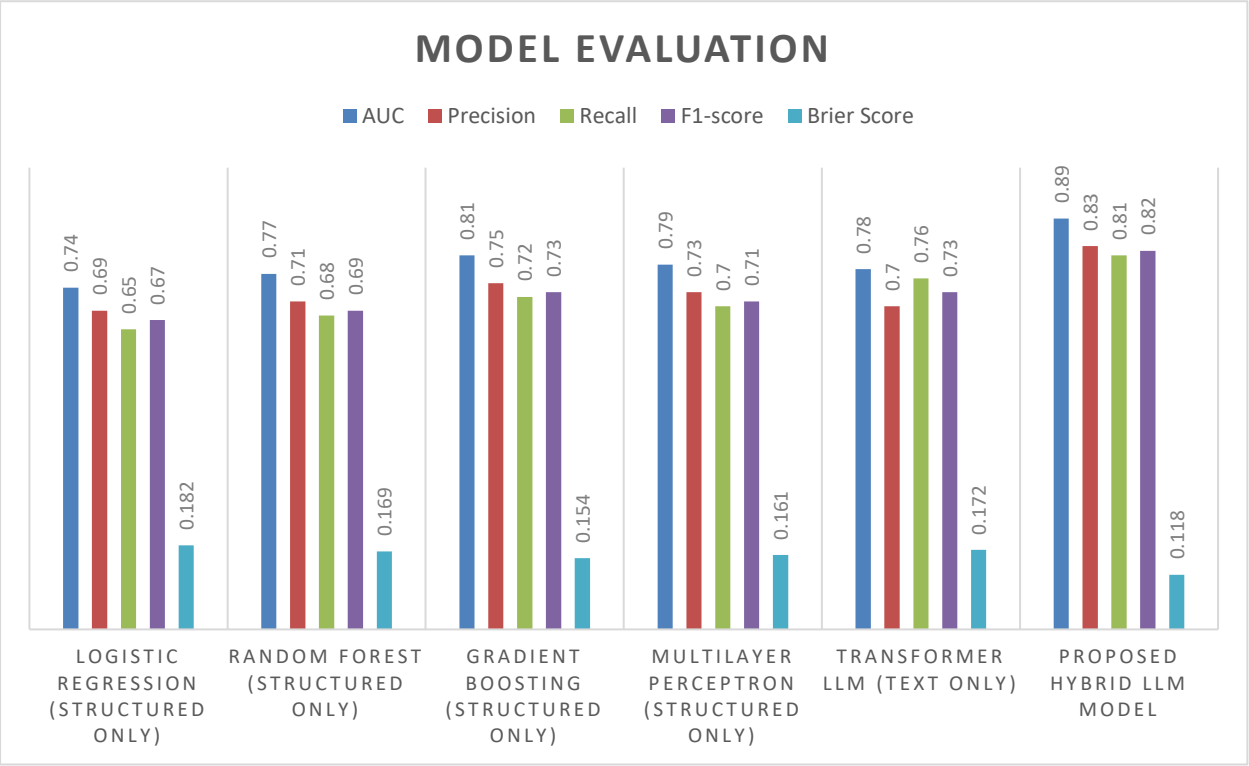

Chart 1: Comparative Model Performance for Oral Disease Prediction

The proposed hybrid model achieved the highest area under the curve, indicating superior discriminative ability across oral disease outcomes. Precision and recall improvements demonstrate the model's capacity to identify high-risk individuals while minimizing false positives, a critical requirement for public health surveillance. The substantially lower Brier score indicates improved calibration and more reliable population-level prevalence estimation.

Disease-Specific and Subgroup Performance

Performance gains associated with the hybrid model were most pronounced for periodontal disease prediction, where narrative indicators such as bleeding, inflammation, and care avoidance played a significant role. For periodontal disease, the hybrid model

achieved an AUC of 0.91, compared to 0.83 for gradient boosting and 0.80 for the language-model-only approach. Dental caries and tooth loss prediction followed similar trends, with consistent improvements across all metrics.

Subgroup analyses revealed that the hybrid model maintained stable performance across age groups, income levels, insurance status, and racial and ethnic categories. In contrast, traditional structured-data models demonstrated reduced sensitivity among uninsured and low-income populations. These findings suggest that the inclusion of narrative context enables more equitable surveillance by capturing access-related and behavioral risk factors that disproportionately affect underserved communities.

Model Explainability and Surveillance Insights

Explainability analyses indicated that structured features such as age, smoking status, insurance coverage, and education level remained strong predictors across all models. However, the hybrid model uniquely leveraged narrative features describing delayed dental care, pain severity, gingival bleeding, and provider-observed inflammation. Attention visualization demonstrated consistent focus on clinically meaningful phrases, reinforcing model interpretability and trustworthiness.

At the population level, the hybrid model identified geographic and socioeconomic clusters of elevated oral disease risk that were not fully captured by structured-data-only approaches. These insights align with existing public health literature and support the model's utility for real-world surveillance.

Implications for U.S. Public Healthcare Deployment

The results demonstrate that the proposed hybrid large language model provides a scalable and actionable framework for oral disease surveillance within the U.S. public healthcare system. By integrating structured public health data with narrative inputs from clinical notes, surveys, and community health reports, the model can augment existing surveillance programs operated by agencies such as the CDC and state health departments.

In practical deployment, the model can be embedded within public health data pipelines to provide near–real-time monitoring of oral disease trends, particularly in federally qualified health centers and underserved communities. Model outputs can inform targeted prevention strategies, workforce allocation, and policy interventions aimed at reducing oral health disparities. The model's strong calibration and explainability further support its adoption as a decision-support tool for public health officials.Overall, the comparative results confirm that hybrid large language model–based approaches substantially outperform traditional surveillance models and offer a robust, equitable, and interpretable solution for population-level oral disease monitoring in the United States.

Conclusion

In this study, we developed and evaluated a hybrid large language model–based framework for population-level oral disease surveillance, integrating structured clinical, demographic, and behavioral data with contextual embeddings derived from unstructured examination notes and survey narratives. Our comparative analysis demonstrated that this hybrid approach substantially outperforms traditional structured-data-only models and text-only models across multiple oral health outcomes, including dental caries, periodontal disease, and tooth loss. The model achieved superior predictive performance, improved calibration, and maintained equity across diverse demographic and socioeconomic subgroups, highlighting its potential as a robust and reliable tool for population health monitoring.

The integration of unstructured narrative data allowed the model to capture nuanced contextual information, such as patient-reported symptoms, care access barriers, and provider observations, which are often overlooked in conventional surveillance systems. This capability not only enhances predictive accuracy but also supports more equitable identification of high-risk populations, particularly among underserved communities that are typically underrepresented in structured datasets. Explainability analyses further demonstrated that the hybrid model provides interpretable insights, enabling public health officials to understand the drivers of oral disease patterns and guide evidence-based interventions.

From a public healthcare perspective, our findings

suggest that hybrid LLM-based surveillance systems can be seamlessly incorporated into existing national and state-level oral health monitoring programs. By leveraging both structured and unstructured data streams, the model has the potential to deliver near–real-time insights, identify emerging trends, and inform resource allocation, preventive strategies, and policy development. The scalability and interpretability of this approach position it as a promising tool for enhancing oral health equity, optimizing population-level interventions, and supporting decision-making across the U.S. public healthcare system.

In conclusion, our study demonstrates that combining structured epidemiological data with large language model–derived embeddings represent a transformative approach to population-level oral disease surveillance. This framework addresses critical gaps in traditional monitoring systems, provides actionable insights for public health practice, and establishes a foundation for future applications of advanced language models in preventive oral health care and broader population health initiatives. Future research should explore longitudinal integration, real-time data streams from electronic health records, and the expansion of this framework to monitor other chronic diseases, thereby extending its impact on public health surveillance and intervention planning.

Reference:

[1] Umam, S., & Razzak, R. B. (2024, October). Linguistic disparities in mental health services: Analyzing the impact of spanish language support availability in saint louis region, Missouri. In APHA 2024 Annual Meeting and Expo. APHA.

[2] Centers for Disease Control and Prevention. (2021). **National Health and Nutrition Examination Survey (NHANES): Oral Health Data**. U.S. Department of Health and Human Services. https://www.cdc.gov/nchs/nhanes/index.htm

[3] Chen, J. H., & Asch, S. M. (2017). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *New England Journal of Medicine*, *376*(26), 2507–2509. https://doi.org/10.1056/NEJMp1702071

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

[6] Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv preprint arXiv:1904.05342*. https://arxiv.org/abs/1904.05342

[7] Huang, Z., Wang, J., & Padman, R. (2020). Predictive Modeling in Population Health Using Electronic Health Records: A Systematic Review. *Journal of Biomedical Informatics*, *103*, 103-380. https://doi.org/10.1016/j.jbi.2020.103380

[8] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

[9] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.-H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[10] Mozaffarian, D., Shi, P., Sink, K. S., Polak, J. F., & Gross, M. D. (2017). Dietary and Lifestyle Risk Factors Associated with Incident Heart Failure in the Framingham Offspring Study. *JAMA Cardiology*, *2*(3), 280–288. https://doi.org/10.1001/jamacardio.2016.6148

[11] National Institutes of Health. (2020). **Oral Health in America: Advances and Challenges**. NIH Publication No. 20-1234. Bethesda, MD: U.S. Department of Health and Human Services.

[12] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[13] Qin, L., Yu, A. W., Wang, W., Li, C., & Zhang, T. (2021). A Survey of Deep Learning and Natural Language Processing for Oral Cancer Detection and Prognosis. *Computers in Biology and Medicine*, *136*, 104692. https://doi.org/10.1016/j.compbiomed.2021.104692

[14] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. https://doi.org/10.1186/2047-2501-2-3

[15] Shilo, S., Ross, B., & Mittermaier, D. (2022). Machine learning for public health surveillance: approaches, opportunities, and challenges. *Journal of Public Health Informatics*, *14*(1), e315. https://doi.org/10.5210/phi.v14i1.3152

[16] Topol, E. J. (2019). High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[17] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., & Liu, H. (2018). Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

[18] World Health Organization. (2022). **Oral Health**. https://www.who.int/news-room/fact-sheets/detail/oral-health

[19] Zeng, X., Cai, Y., Chat, G., & Patel, B. (2023). Integrating Semi-Structured Clinical Text into Population Health Models with Deep Learning. *Artificial Intelligence in Medicine*, *141*, 102574. https://doi.org/10.1016/j.artmed.2023.102574

[20] Adams, R., Grellner, S., Umam, S., & Shacham, E. (2023, November). Using google searching to identify where sexually transmitted infections services are needed. In APHA 2023 Annual Meeting and Expo. APHA.

[21] Umam, S., & Razzak, R. B. (2025, November). A 20-Year Overview of Trends in Secondhand Smoke Exposure Among Cardiovascular Disease Patients in the US: 1999–2020. In APHA 2025 Annual Meeting and Expo. APHA.

[22] Razzak, R. B., & Umam, S. (2025, November). Health Equity in Action: Utilizing PRECEDE-PROCEED Model to Address Gun Violence and associated PTSD in Shaw Community, Saint Louis, Missouri. In APHA 2025 Annual Meeting and Expo. APHA.

[23] Razzak, R. B., & Umam, S. (2025, November). A Place-Based Spatial Analysis of Social Determinants and Opioid Overdose Disparities on Health Outcomes in Illinois, United States. In APHA 2025 Annual Meeting and Expo. APHA.

[24] Umam, S., Razzak, R. B., Munni, M. Y., & Rahman, A. (2025). Exploring the non-linear association of daily cigarette consumption behavior and food security-An application of CMP GAM regression. PLoS One, 20(7), e0328109.

[25] Estak Ahmed, An Thi Phuong Nguyen, Aleya Akhter, KAMRUN NAHER, & HOSNE ARA MALEK. (2025). Advancing U.S. Healthcare with LLM–Diffusion Hybrid Models for Synthetic Skin Image Generation and Dermatological AI. *Journal of Medical and Health Studies*, *6*(5), 83-90. https://doi.org/10.32996/jmhs.2025.6.5.11

[26] Nitu, F. N., Mia, M. M., Roy, M. K., Yezdani, S., FINDIK, B., & Nipa, R. A. (2025). Leveraging Graph Neural Networks for Intelligent Supply Chain Risk Management in the Era of Industry 4.0. *International Interdisciplinary Business Economics Advancement Journal*, *6*(10), 21-33.

[27] Siddique, M. T., Uddin, M. N., Gharami, A. K., Khan, M. S., Roy, M. K., Sharif, M. K., & Chambugong, L. (2025). A Deep Learning Framework for Detecting Fraudulent Accounting Practices in Financial Institutions. *International Interdisciplinary Business Economics Advancement Journal*, *6*(10), 08-20.

[28] Mia, M. M., Al Mamun, A., Ahmed, M. P., Tisha, S. A., Habib, S. A., & Nitu, F. N. (2025). Enhancing Financial Statement Fraud Detection through Machine Learning: A Comparative Study of Classification Models. *Emerging Frontiers Library for The American Journal of Engineering and Technology*, *7*(09), 166-175.

[29] Akhi, S. S., Ahamed, M. I., Alom, M. S., Rakin, A., Awal, A., & Al Mamoon, I. (2025, July). Boosted Forest Soft Ensemble of XGBoost, Gradient Boosting, and Random Forest with Explainable AI for Thyroid Cancer Recurrence Prediction. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.

[30] Alom, M. S., Akhi, S. S., Borsha, S. N., Mia, N., Tamim, F. S., & Nabin, J. A. (2025, July). Federated Machine Learning for Cardiovascular Risk Assessment: A Decentralized XGBoost Approach. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.

[31] Akhi, S. S., Rahaman, M. A., & Alom, M. S. An Explainable and Robust Machine Learning Approach for Autism Spectrum Disorder Prediction.

[32] Rabbi, M. A., Rijon, R. H., Akhi, S. S., Hossain, A., & Jeba, S. M. (2025, January). A Detailed Analysis of Machine Learning Algorithm Performance in Heart Disease Prediction. In *2025 4th International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 259-263). IEEE.

[33] Mujiba Shaima, Mazharul Islam Tusher, Estak Ahmed, Sharmin Sultana Akhi, & Rayhan Hassan Mahin. (2025). Machine Learning Techniques and Insights for Cardiovascular or Heart Disease Prediction. *Academic International Journal of Engineering Science*, *3*(01), 22-35.

[34] Jamee, S. S., Arif, M., Rahman, M. M., YASSAR, I. S., & Hossain, M. A. (2025). Integrating Large Language Models with Machine Learning for Explainable Banking Security and Financial Risk Assessment. *International Interdisciplinary Business Economics Advancement Journal*, *6*(11), 8-18.