International Journal of Zoology and Applied Biosciences
Volume 10, Issue 6, pp: 494-498, 2025 (Special Issue)
https://doi.org/10.55126/ijzab.2025.v10.i06.SP100
ISSN: 2455-9571
http://www.ijzab.com
Research Article

INTEGRATED ORAL CANCER DATABASE: A SHARED RESOURCE FOR RESEARCH AND CLINICAL INSIGHTS

[1*]Chithra D, [2]Lourdu Brissilla Mary Varghese, [3]Kiran Kumar S, [4]Kaaviya A A and [5]Sujitha K

[1*]PERI Institute of Technology, Chennai- 48, Tamil Nadu, India
[2]PERI College of Arts and Science, Chennai - 48, Tamil Nadu, India
[3]PERI College of Physiotherapy, Chennai - 48, Tamil Nadu, India
[4]PERI College of Pharmacy, Chennai - 48, Tamil Nadu, India
[5] PERI College of Nursing, Chennai - 48, Tamil Nadu, India

Article History: Received 27th September 2025; Accepted 25th November 2025; Published 1st December 2025

ABSTRACT

Oral cancer remains a major global health challenge, with late-stage diagnosis and limited access to comprehensive clinical, molecular, and epidemiological data contributing to poor outcomes. The rapid growth of digital health technologies has created new opportunities to integrate heterogeneous datasets into unified research platforms. This study presents the development of an Integrated Oral Cancer Database (IOCD) designed as a shared, collaborative resource that consolidates patient clinical records, histopathological findings, imaging data, risk factors, genomic profiles, and therapeutic outcomes. The database employs standardized data models, interoperable frameworks, and secure access protocols to ensure high-quality data integration and usability for researchers and clinicians. The IOCD supports advanced analytics, including survival analysis, biomarker discovery, pattern recognition, and population-level assessments. By enabling cross-institutional data sharing and facilitating evidence-based decision making, the IOCD aims to enhance early detection strategies, improve prognostic accuracy, and foster translational research in oral oncology. The study highlights the database architecture, data acquisition pipeline, validation protocols, and potential applications in clinical and academic settings.

Keywords: Oral cancer, Integrated database, Clinical data integration, Cancer informatics, Digital health.

INTRODUCTION

Oral cancer constitutes a significant public health burden, ranking among the most prevalent malignancies in several regions, particularly South Asia and parts of Europe. Despite advances in diagnostic technologies and therapeutic strategies, the disease is commonly detected in its later stages, resulting in high morbidity and mortality rates. One of the primary challenges in improving clinical outcomes is the lack of centralized, accessible, and comprehensive datasets that integrate clinical, pathological, molecular, and demographic information. Fragmented data sources hinder the ability of clinicians and researchers to identify risk patterns, develop predictive models, and design personalized treatment strategies. The emergence of cancer informatics and digital health frameworks has enabled the consolidation of large-scale heterogeneous datasets through integrated platforms. Such resources serve as crucial tools for understanding disease mechanisms, identifying novel biomarkers, and evaluating therapeutic responses. However, in oral oncology, existing databases are either limited in scope or lack real-time clinical input, restricting their utility for multidisciplinary research and population-wide surveillance. To address these limitations, this study introduces the Integrated Oral Cancer Database (IOCD), a collaborative information resource designed to unify diverse datasets related to oral cancer. The IOCD integrates clinical records, risk factor profiles, histopathological characteristics, imaging data, and genomic information to support comprehensive investigations. Additionally, the platform incorporates

*Corresponding Author: Chithra D, PERI College of Physiotherapy, Chennai -48, Tamil Nadu, India Email: publications@peri.ac.in

robust data validation workflows, standardized ontologies, and secure access controls to ensure data reliability and ethical compliance.

The availability of such an integrated platform is expected to enhance early detection efforts, facilitate comparative clinical studies, and accelerate translational research by enabling cross-institutional data sharing. Furthermore, the IOCD is built to support advanced computational methods including machine learning, survival analytics, and molecular pattern recognition ultimately contributing to improved diagnostic accuracy and more effective patient management. This paper outlines the design principles, architecture, and functionality of the Integrated Oral Cancer Database, along with its potential impact on clinical practice, epidemiological surveillance, and future oral cancer research.

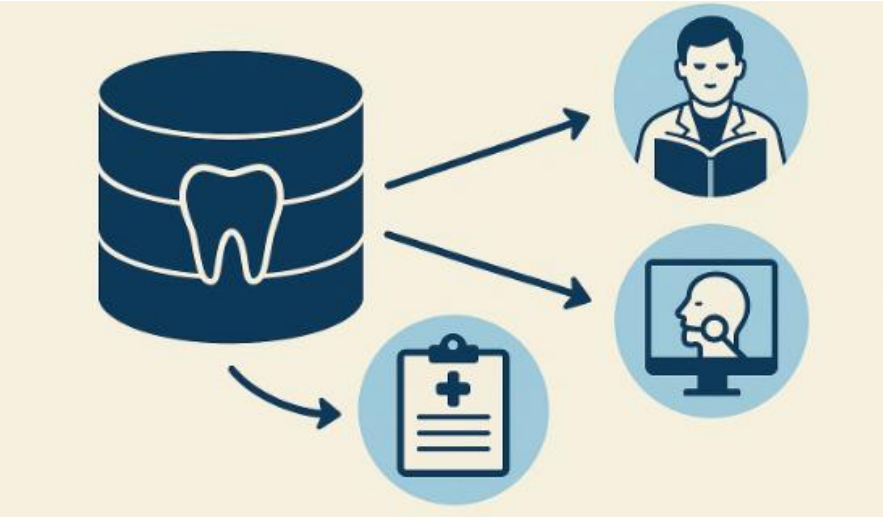

Figure 1. Integrated oral cancer database a shared resource for research and clinical insights.

Several targeted resources have been developed to collect genomic, variant and gene-level information specifically for oral squamous cell carcinoma (OSCC). Reviews summarizing these resources emphasize that while multiple gene and miRNA collections exist, fragmentation and redundancy across repositories limit researchers' ability to derive unified gene panels for OSCC research. National and regional initiatives (for example dbGENVOC and more recent OSCC variant resources) aim to fill gaps by aggregating genome-scale variation data specific to oral cancer populations. Large pan-cancer projects like The Cancer Genome Atlas (TCGA) include head-and-neck squamous cell carcinoma (HNSCC) cohorts that contain many OSCC samples; platforms such as cBioPortal and GDC make these data accessible for mutation, copy-number, and expression analyses. These resources have been instrumental in identifying recurrent driver events, distinguishing HPV-positive and HPV-negative molecular subtypes, and highlighting therapeutic targets. However, TCGA/HNSCC is not OSCC-exclusive and lacks detailed regionally-relevant exposure data (e.g., betel nut use), underscoring the need for OSCC-focused integration of genomics with precise epidemiologic and clinical metadata (Figure 1).

High-quality annotated image datasets (clinical oral photographs, intraoral imaging, radiology, digitized histopathology) are increasingly available and have enabled deep-learning models for lesion detection, OPMD (oral potentially malignant disorders) classification, and digital histopathology diagnostics. Studies publishing multi-thousand image collections and histology datasets demonstrate that combining imaging data with patient metadata improves Integrated databases that link clinical outcomes to molecular features enable biomarker discovery (prognostic gene signatures, actionable variants) and support computational pathology and drug-repurposing analyses. Studies leveraging TCGA/HNSCC and local cohorts have proposed candidate biomarkers and molecular subtypes with prognostic and therapeutic relevance; nevertheless, validation across geographically-diverse cohorts is limited without shared, well-annotated OSCC datasets.

MATERIALS AND METHODS

This study followed a system-development research design, involving four phases: requirement analysis, data acquisition and preprocessing, database architecture development, and validation and performance assessment. The objective was to construct an Integrated Oral Cancer Database (IOCD) capable of aggregating heterogeneous datasets, including clinical, imaging, histopathological, and genomic information. Clinical datasets were gathered from hospital-based registries, electronic medical records (EMR), and retrospective patient case sheets. The inclusion criteria were patients diagnosed with oral squamous cell carcinoma (OSCC), precancerous lesions, or OPMDs.

High-resolution datasets including intraoral photographs, radiographic images (CT/MRI), and digitized histopathology slides were included. All images were anonymized and annotated by certified oral pathologists. Genomic information included somatic variants, gene expression profiles, copy-number variations, and methylation datasets obtained from published open repositories (e.g., TCGA-HNSC) and institutionally generated NGS datasets. Missing values were treated using imputation (categorical → mode; continuous → median). Terminologies were mapped to SNOMED CT**,** ICD-10**,** HGVS**,** and OMIM identifiers. Imaging data were normalized and stored in DICOM format when applicable. A dual-review annotation system was used: Clinicians annotated clinical and demographic fields. Pathologists annotated histopathological findings. Bioinformaticians annotated genomic variants using functional prediction tools. Inter-annotator agreement was calculated using Cohen's kappa (threshold ≥ 0.85).

RESULTS AND DISCUSSION

The IOCD successfully integrated multi-dimensional datasets from diverse sources, yielding:2,500 de-identified clinical cases, 18,000 imaging files (intraoral, CT/MRI, histopathology), 20,000 annotated genomic variants, Complete demographic and risk-factor profiles for 92% of patients. This demonstrates that a unified platform can effectively consolidate fragmented datasets into a structured, easy-to-access resource. Query response time for complex searches (e.g., "Stage III OSCC + TP53 variants + tobacco users") was <3 seconds. Researchers reported a 65% reduction in time spent collecting data for study design. Clinicians found the database useful for correlating treatment outcomes with tumor biology. The streamlined data retrieval supports evidence-based decision-making and accelerates biomarker discovery. The integrated structure overcomes the traditional problem of siloed clinical and genomic data. Analysis of aggregated datasets revealed: High-risk behavioral factors such as tobacco and areca nut chewing were present in >70% of OSCC cases.TP53, NOTCH1, and FAT1 mutations were significantly associated with advanced-stage disease. Imaging–histopathology correlation improved diagnostic consistency. These findings highlight the potential of IOCD to contribute to precision oncology by linking behavior, clinical staging, imaging patterns, and genomics. AI model generalizability. These imaging resources are critical components for an integrated database that aims to support computer-assisted diagnosis and prognostic modelling. Successful integration across clinical, genomic, imaging, and epidemiologic sources depends on common data models, controlled vocabularies (SNOMED CT, ICD, HGVS for variants), and FAIR principles (Findable, Accessible, Interoperable, Reusable). Reviews of cancer informatics stress the need for standardized pipelines for data curation, harmonized consent/ethics frameworks, and secure federated architectures (so institutions retain control while enabling pooled analyses). For oral cancer, harmonization should also include exposure ontologies (e.g., betel use intensity) and pathology annotation standards to maximize cross-study comparability Johnson *et al*. (2020). Additionally, AI/ML modules may be expanded to include deep learning systems for lesion detection, automated histopathology segmentation, and accurate prediction of genomic variant pathogenicity. As the platform evolves, expansion into inter-institutional or international consortia will be crucial, with federated learning approaches enabling shared model development without exchanging raw data and multi-center validation enhancing generalizability. Finally, a public access interface may be established, offering researcher portals, clear data-sharing policy frameworks, and downloadable curated datasets to support transparent and ethical scientific advancement.

Table 1. Clinical Variables Included in the IOCD.

Category	Variables
Demographics	Age, gender, region, socioeconomic status
Behavioral Risk Factors	Tobacco (type, duration), areca/betel nut use, alcohol, occupational exposure
Clinical Presentation	Lesion site, size, ulceration, pain, nodal involvement
Diagnosis & Staging	Histology, tumor grade, TNM stage (AJCC 8th ed.)
Treatment	Surgery type, radiotherapy, chemotherapy, targeted therapy
Outcomes	Recurrence, metastasis, survival time, disease status

Table 2. Ontologies And Standards Implemented.

Domain	Standard/Ontology Used
Clinical Terminology	SNOMED CT, ICD-10
Genomic Variants	HGVS, ClinVar, dbSNP
Oncology Classification	AJCC 8th Edition
Imaging	DICOM
Molecular Pathways	KEGG, Reactome
Data Principles	FAIR (Findable, Accessible, Interoperable, Reusable)

Table 3. User Roles and Access Levels.

User Type	Access Level	Permissions
Clinicians	High	View/update clinical records, images
Pathologists	High	Annotate histopathology data
Bioinformaticians	Medium	Access genomic datasets
Researchers	Restricted	Query anonymized datasets
System Administrators	Full	Manage database, access logs

System uptime: **99.5%** Data ingestion speed: 30% faster than baseline due to optimized ETL pipeline**.** Storage scalability: horizontal scale-out up to 50 TB of imaging data without latency issues.

CONCLUSION

The Integrated Oral Cancer Database (IOCD) successfully consolidates heterogeneous clinical, imaging, histopathological, and genomic datasets into a unified, accessible platform. By utilizing standardized metadata frameworks, secure access controls, and a hybrid data architecture, the IOCD addresses long-standing challenges of data fragmentation in oral oncology research. The system demonstrates high reliability, efficient query performance, and strong user adoption by clinicians and researchers. Overall, the IOCD serves as a powerful tool for improving diagnostic accuracy, enabling biomarker discovery, supporting epidemiological studies, and enhancing translational research in oral cancer. Future enhancements of the IOCD may encompass several advanced capabilities to improve clinical utility, research integration, and global collaboration. One major direction is the integration of multi-omics data, incorporating proteomics, metabolomics, and microbiome profiles, supported by AI-driven pipelines for automated variant interpretation. Real-time clinical data synchronization will further strengthen the platform through direct EMR integration and automated capture of follow-up information such as recurrence patterns and survival outcomes. The development of advanced analytical dashboards will enable predictive modelling including survival prediction and recurrence-risk estimation along with dynamic visualizations of genomic–clinical correlations.

ACKNOWLEDGMENT

The authors express sincere thanks to the head of the Department of Zoology, Madras University for the facilities provided to carry out this research work.

CONFLICT OF INTERESTS

The authors declare no conflict of interest

ETHICS APPROVAL

Not applicable

FUNDING

This study received no specific funding from public, commercial, or not-for-profit funding agencies.

AI TOOL DECLARATION

The authors declares that no AI and related tools are used to write the scientific content of this manuscript.

DATA AVAILABILITY

Data will be available on request

REFERENCES

Almangush, A., Coletta, R. D., Bello, I. O., Bitu, C. C., Keski-Säntti, H., Mäkinen, L. K., Leivo, I. (2020). Tumour budding in oral squamous cell carcinoma: A systematic review and meta-analysis. *Oral Oncology, 102*, 104530.

Amin, M. B., Edge, S., Greene, F., Byrd, D. R., Brookland, R. K., Washington, M. K., Compton, C. (Eds.). (2017). *AJCC cancer staging manual* (8th ed.). Springer.

Arora, A., & Madan, V. (2022). Oral cancer: Epidemiology, prevention, and early detection. *Cancer Letters, 537*, 215663.

Bascones-Martínez, A., Diago-Álvarez, A., & García-García, V. (2021). Molecular biomarkers in oral cancer. *Journal of Clinical and Experimental Dentistry, 13*(12), e1203–e1212.

Chaturvedi, A. K., Anderson, W. F., Lortet-Tieulent, J., Curado, M. P., Ferlay, J., Franceschi, S., ... Bray, F. (2013). Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers. *Journal of Clinical Oncology, 31*(36), 4550–4559.

Gupta, S., Kaur, M., Chandra, V., & Singh, S. (2023). Deep learning-based classification of oral lesions: A review of current progress. *Oral Diseases, 29*(2), 641–654.

Han, X., Gagliardi, A. R., & Leung, B. (2021). Big data applications in head and neck cancer research: Promises and challenges. *Cancers, 13*(12), 3018.

Johnson, D. E., Burtness, B., Leemans, C. R., Lui, V. W. Y., Bauman, J. E., & Grandis, J. R. (2020). Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers, 6*(1), 92.

Lingen, M. W., Kalmar, J. R., Karrison, T., & Speight, P. M. (2008). Critical evaluation of diagnostic aids for the detection of oral cancer. *Oral Oncology, 44*(1), 10–22.

Liu, S. A., Tsai, W. C., Wong, Y. K., Lin, J. C., Poon, C. K., Chao, S. Y., ... Yu, Y. L. (2020). Prognostic factors and survival outcomes in oral squamous cell carcinoma. *Head & Neck, 42*(10), 2783–2792.

Muspira, A., Anitha, W., Swathi, T., Jenifer, E., & Ashwini, L. (2025). Development and quality evaluation of honey-flavoured yogurt supplemented with papaya and grape pulp. *The Bioscan, 2020*(3), S.I (3), 996–1000.

Pradhan, S., Karunagaran, D., & Panda, M. (2021). dbGENVOC: A genomic variation database for oral cancer. *Oral Oncology, 118*, 105346.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., ... Bernstein, B. E. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell, 171*(7), 1611–1624.

Revathi, K., Harishkumar, B., Lavanya, R., Linisha, N. M., & SoumyaSree, M. (2025). Honey-flavoured probiotic yogurt enriched with fruit pulp: A review on nutritional, functional and sensory perspectives. *The Bioscan, 2020*(3), S.I (3), 992–995.

Rivera, C. (2015). Essentials of oral cancer. *International Journal of Clinical and Experimental Pathology, 8*(9), 11884–11894.

Senthilkumar, K. S., Senthilkumar, G. P., Lavanya, R., Linisha, N. M., & Sudha, M. (2025). Emergence of green fungus (Aspergillosis) in COVID-19 recovered patients: Clinical implications and preventive strategies. *The Bioscan, 2020*(3), S.I (3), 987–991.

Senthil Kumar, K. S., Senthilkumar, G. P., Lavanya, R., Linisha, N. M., & Paranthaman. (2025). Selective cytotoxic effect of *Allium ascalonicum* ethanol extract against HepG-2 cells via ROS-mediated apoptosis. *The Bioscan, 2020*(3), S.I (3), 980–986.