FaceBase
A Resource For Craniofacial Researchers

2025 FaceBase Community Forum - Meeting Summary

June 3 & 4, 2025
USC's Information Sciences Institute

Executive Summary

Day One began with introductions and updates from Jennifer Webster-Cyriaque, Debara Tucci, Susan Gregurick ((acting) directors of NIDCR, NIDCD, and ODSS, respectively) and Yang Chai (co-PI of FaceBase) regarding FaceBase's evolution and its strategic partnerships, particularly focusing on data-driven research and collaboration with ODSS, NIDCR and NIDCD. The FaceBase team provided an overview of progress made in the first year of the current funding cycle and outlined a strategic road map for the upcoming four years. Keynote speaker, Helen Berman, discussed the history and evolution of the Protein Data Bank. Subsequent sessions highlighted recent advances in identifying genetic factors underlying craniofacial birth defects, and the application of avian embryos as research models in craniofacial research. The latter part of the day addressed technical and clinical aspects, featuring panel discussions on integrating clinical components into FaceBase, developing medical applications for remote support, and analyzing large-scale datasets to enhance diagnostic precision and patient outcomes.

Day Two of the meeting emphasized the advancement of dental, oral, and craniofacial research by leveraging FaceBase data collections and sharing. Discussions centered on data reuse, effective data science strategies, and community engagement. The keynote speaker, Axel Visel, highlighted the importance of engaging and educating the community to optimize data collection and data sharing practices, presenting insights from the NIDCR Data Science Strategies working group. Key topics included integrating diverse data sources, improving data accessibility, and fostering innovation. Dr. Visel also addressed the challenges associated with managing and utilizing large-scale datasets, highlighting the necessity of collaborative approaches to accelerate research progress and clinical translation.

Researchers presented findings demonstrating how FaceBase data has been invaluable to study variations in facial morphology and inform respiratory mask design specifically for the pediatric community. Their work highlights the critical need for demographic-related facial variations in medical device development and clinical practice. Additionally, as part of the development of interdisciplinary work with NIDCD, a presentation explored macroglossia, a

condition characterized by tongue enlargement, and detailed how integrating clinical, imaging and multi-omics datasets from Beckwith-Wiedemann Syndrome cohorts within FaceBase will improve diagnostic accuracy and inform improved clinical guidelines.

A final panel discussed the challenges and opportunities in integrating diverse data types into platforms such as FaceBase. The panel emphasized the need for standardized data models, clearly defined data standards, and community input in shaping data quality and integration protocols.

In addition to the rich scientific program, the Forum marked several milestones for FaceBase. We celebrated receipt of the CoreTrustSeal accreditation; were recognized as one of a select number of NIH-approved Controlled Access Data Repositories (CADRs) qualified in handling sensitive genomic and other human-derived data; were designated an NIH HEAL-approved repository, launched a new seq-FISH Spatial Genomics data type; and entered formal collaborations with NIDCD (EarBase temporal bone archive) and the CranioRate federation. The achievement of these milestones mark strong progress along the FaceBase 4 roadmap: TRUST/FAIR compliance, AI-ready data, community engagement, and interoperability with LLMs and notebooks.

Day One

Introductions by Institute Directors

Dr. Jennifer Webster-Cyriaque, acting director of NIDCR, emphasized the importance of data science in advancing oral and overall health and NIDCR's commitment to aligning with the Real World Data Initiative. To that end, Webster-Cyriaque described FaceBase as an important "hybrid" repository that provides the missing link between specialized repositories (e.g., GEO) and generalist repositories (e.g., figshare). Dr. Webster-Cyriaque discussed FaceBase's key role in NIH data science strategies, including the end goal of federating high-quality data to help answer scientific questions holistically. Together with the focus on AI-readiness, FAIRness and TRUSTworthiness, FaceBase serves as an important bridge between common data elements to data analysis and data generation.

Dr. Debara Tucci, director of the National Institute on Deafness and Other Communication Disorders (NIDCD), provided an overview of the Institute's mission areas, and the importance of obtaining human tissues for making discoveries that lead to cures. Dr. Tucci described the ongoing impact of human temporal bone research, highlighting the utility of combining tissue-level analysis with high-resolution imaging and the medical record. She also shared an overview of NIDCD's Temporal Bone Initiative, which combines a nationwide hearing and balance donor

program, a cooperative network of affiliated research laboratories, and the collaboration with Face Base to redesign the temporal bone database to serve as a repository and hub for tissue sharing and advanced analysis. In that context, Dr. Tucci also highlighted how the collaboration with FaceBase and NIDCR will explore overlaps in temporal bone structure and craniofacial structure and further elucidate the connection between diseases and syndromes.

Dr. Susan Gregurick, Associate Director for Data Science and Director of the Office of Data Science Strategy (ODSS) at the National Institutes of Health (NIH), discussed ODSS's commitment to sustaining NIH-funded research and strengthening the NIH Data Repository and Knowledgebase landscape. She highlighted how FaceBase's contributions to developing and following best practices in all aspects of the data value chain (i.e. data creation, data transformation, data value-add, and data impact), along with its recent CoreTrustSeal accreditation, aid in increasing the return on investment of NIH-funded research.

FaceBase Updates

Drs. Carl Kesselman and Rob Schuler (Information Sciences Institute, USC) presented on FaceBase's updates and achievements, highlighting its expansion to biologically/anatomically relevant domains and the development of new strategic partnerships to further drive this expansion. New collaborations include the partnership with NIDCD, the craniosynostosis patient data federation project with CranioRate), and approval as an NIH HEAL Initiative repository. A new data type has also been released on FaceBase: seq-FISH based Spatial Genomics. FaceBase demonstrated its credibility as a resource for the DOC research community through its CoreTrustSeal accreditation after a two-year approval process as well as becoming one of a select number of NIH approved Controlled Access Data Repositories (CADRs) handling genomic data. This latter effort is part of our ongoing effort to adopt the NIST 800-53 Cybersecurity framework in compliance with the revised NIH Genomic Data Sharing (GDS) policy for 2025. These efforts position FaceBase for effective scale for use to researchers in other anatomically/biologically-relevant health domains.

Schuler walked attendees through the roadmap of FaceBase 4 that includes:

- Addressing community needs and engagement through the full translational spectrum, supporting cross-disciplinary data through outreach and training, and providing resources to help researchers develop their NIH Data Management and Sharing plans and apply best practices in data stewardship.
- Enhancing the quality of services and efficiency of FaceBase operations to serve a growing community through efficient cloud hosting, compliance with TRUST and FAIR principles, development of AI/ML ready data with helpful visualization and analysis tools to promote interoperability with LLMs and notebooks.
- Using best practice governance for data repositories by reviewing policies on access, privacy, and ethics and ensuring we align with users' needs and expectations.

Protein Data Bank's Evolution and Impact

Keynote speaker, Dr. Helen Berman (USC), co-founder of the Protein Data Bank (PDB), discussed the history and evolution of the PDB. Dr. Berman highlighted its role in developing best practices in data management for structural biology. Berman emphasized that the PDB's success relies on continuous innovation of science, technology, and community involvement. The scientific community informed the development of the PDB's practices for data curation, validation, and representation. She also touched on the PDB's impact on fields like structural bioinformatics and drug discovery, and its role in enabling recent advancements in protein structure prediction. This presentation emphasized many strategies that are the founding principles of FaceBase (e.g., importance of data quality and controlled, machine-readable vocabularies and standards). While FaceBase has built on the strategies pioneered by PDB, PDB developed with one central data focus, which ensured a highly specialized and narrow focus. This presents an important consideration for FaceBase as we aim to support the needs of a varied and growing community of researchers.

Craniofacial Birth Defects: Genetic Insights

Dr. Justin Cotney (Children's Hospital of Philadelphia) presented research on superenhancers (non-coding regulatory sequences) and their roles as genetic factors underlying causes of craniofacial birth defects. This work has resulted in the identification of over 100,000 regulatory sequences across the genome and demonstrated examples of the role of these sequences in conditions like oral facial clefting and craniosynostosis. The data from this research will be made available via FaceBase and has the potential for clinical application in developing predictive tools for non-coding variants for use by clinical geneticists.

Avian Embryos in Craniofacial Research

Dr. Samantha Brugmann (Cincinnati Children's Hospital) presented on the use of avian embryos as a model for studying craniofacial development and disease in humans, highlighting their cost-effectiveness, ease of use, and genetic similarity to mammals. She discussed two naturally occurring avian mutants: Talpid 2, which exhibits a range of ciliopathy-related phenotypes including cleft palate and polydactyly, and Cleft Primary Palate (CPP), which shows severe upper beak defects. Dr. Brugmann's research has identified mechanisms linking skeletal development and kidney function in Talpid 2, and she is exploring potential drug treatments for these defects. She emphasized the potential of avian models to provide insights into human diseases and the need for further genomic comparisons between species to enhance their utility in research.

Both Dr. Cotney and Dr. Brugmann's presentations highlight the power of having diverse model organisms to support the study of craniofacial development and FaceBase's utility as a resource for craniofacial researchers for data reuse.

Clinical Data Integration in FaceBase (Panel)

The day's panel, "Incorporating Relevant Clinical Elements (EMR/EDR) into FaceBase to Effectively Facilitate Clinical Research", formed by Drs. Parish Sedghizadeh, Mohammad Khalifeh and Anette Vistoso (Herman Ostrow School of Dentistry/USC), discussed incorporating clinical elements and data into FaceBase, focusing on specific conditions and diseases with available patient data. They emphasized the importance of specific structured clinical data elements to enable effective research and improve clinical decision-making.

Dr. Vistoso described the use of a highly specific and user-friendly platform to collect structured clinical data to feed into AI/ML models for effective precision medicine. She proposed improving real-time diagnoses by connecting repositories with point-of-care facilities to provide structured notes, information on craniofacial biomarkers, and patient-reported outcomes.

Dr. Sedghizadeh described progress on elements of a new pilot project to integrate clinical elements from patients with temporomandibular disorders (TMD) at USC into FaceBase – the most crucial of which is the diagnosis with ICD-10 codes. Sedghizadeh highlighted the need for clear patient consent regarding data use in repositories (that specifically calls out use of identifiable health information for research without requiring re-authorization) and the potential of AI/ML methods to analyze clinical notes and improve diagnostic accuracy.

Dr. Khalifeh presented a case study using myTMD, an AI-powered mobile application for orofacial pain disorders, which showed promising initial results in a small study, though further research is needed to validate its accuracy across a broader patient population. The application aims to standardize patient questionnaires, increase awareness of conditions, and facilitate coordination between clinicians and patients.

Critical conclusions from the panel discussion include the development of approaches to addressing data management challenges and opportunities for these structured, standardized datasets as well as potential for crowdsourcing clinical data and the importance of funding for such projects.

AI for Multimodal Data Analysis

Dr. Thomas Peterson (UCSF) provided an overview of the NIH HEAL Initiative's efforts to integrate clinical data into research on chronic low back pain. He discussed methods for

analyzing large datasets, particularly focusing on multimodal phenotyping and data integration. He explained how latent class models can be used to identify biologically meaningful subgroups within patient populations, reducing data dimensionality and making it more accessible for AI analysis. The presentation covered statistical modeling approaches, patient-specific outcome prediction, and knowledge integration methods, including the use of clinical data, published literature, and existing datasets.

Peterson also highlighted a tool called Philter that de-identifies clinical notes, making them more accessible for research purposes. He concluded by discussing the potential of AI to improve diagnostic accuracy in temporomandibular disorder (TMD) and orofacial pain through structured note-taking and algorithm-assisted navigation.

Enhancing Diagnostics with AI Systems

Drs. Glenn Clark and Anette Vistoso (USC) focused on improving diagnostic accuracy and patient outcomes through structured data collection and machine learning. Dr. Clark presented MyDocNote, a system using structured notes and algorithmic assistance to aid in diagnosis, highlighting its 93% accuracy in predicting diagnoses from a dataset of 2,000 cases. The system is being expanded to include 10,000 cases and aims to reduce misdiagnosis by collecting comprehensive patient data. The discussion concluded with questions about the system's applicability to telehealth and its potential to refine phenotypes in temporomandibular disorder (TMD) patients.

Dynamic MRI for Craniofacial Imaging

Dr. Krishna Nayak (USC) presented on the potential of dynamic MRI to study temporomandibular disorders (TMD) and other craniofacial and oral conditions. He explained how MRI technology has evolved to become much faster, allowing for dynamic imaging of movements, and discussed the advantages of lower field strength MRI systems for imaging around metal and air. Coupling innovative tools like dynamic MRI with high quality EMR enables powerful new opportunities for research.

Day Two

Meeting DOC Data Infrastructure Challenges

Dr. Axel Visel (Lawrence Berkeley National Lab), presented on the current state and challenges of data management in dental, oral and craniofacial research, highlighting the diversity of data

types and systems used. Key unsolved challenges included the complexity of the data ecosystem, lack of connectivity between systems, and concerns about the sustainability of data repositories.

Four main opportunity areas for applications of data science in support of DOC research identified were:

- Understanding oral health disparities,
- Improving AI and ML readiness of dental data,
- Developing new applications, and
- Leveraging diverse dental data types.

Recommendations were made for NIDCR to invest in robust data infrastructure, integrate across data types, develop better tools, and establish strong data stewardship policies. Visel also pointed to the published report of the NIDCR Data Science Strategy Working Group (https://www.nidcr.nih.gov/sites/default/files/2024-09/nadcrc-dsswg-final-report-may-24.pdf), which provides detailed background information on these challenges, opportunities, and recommendations.

Visel presented the DDS Hub (https://www.ddshub.nih.gov/), a new resource for community engagement and funding opportunities in dental and craniofacial research. The DDS Hub is a central hub for dental data science resources, aiming to support the entire cycle of data-driven research. An important consideration moving forward will be the timeline for addressing AI bias, which is an ongoing challenge that will require continuous adaptation as AI methods evolve.

Facial Shape Variations in Respirator Design

Drs. Christopher Nemeth (Applied Research Associates) and Benedikt Hallgrimsson (University of Calgary) presented their research leveraging data available in FaceBase to study variation in facial morphology to use age- and ancestry-related facial varials to inform the design of pediatric respirators.

Hallgrimsson presented findings on facial shape variation, noting that while age is the primary source of variation, ancestry also plays a significant role, explaining that it accounts for approximately 15% of total facial shape variance. He emphasized the importance of considering both age-related and ancestry-related variations in product design to ensure optimal fit of pediatric respirators. Nemeth discussed how these data can inform respirator design through normalizing baselines, developing CAD models, and evaluating fit across age groups.

Facial Data Analysis Techniques

Dr. Peter Claes (KU Leuven, Belgium) highlighted the use of data available in FaceBase to research clinical applications of 3D facial analysis, emphasizing the importance of 3D imaging in capturing facial data and its potential for use with AI/ML methods. Claes described using statistical and deep learning techniques for analyzing facial data, focusing on both population-level and patient-level hypothesis testing. He demonstrated how facial growth models can be used to create personalized facial signatures and discussed the development of geometric deep learning for 3D facial scans. He also explained how these techniques can be applied to diagnose genetic syndromes and generate synthetic faces for clinical teaching purposes. The talk concluded with a discussion of using facial data to model growth and development across different ages.

Beckwith-Wiedemann Syndrome Research Overview

Dr. Jennifer Kalish (Children's Hospital of Philadelphia) discussed her research on Beckwith-Wiedemann Syndrome (BWS), a rare overgrowth disorder affecting approximately 500 U.S. births annually, with a tenfold increase in incidence in IVF pregnancies. She outlined how BWS research involving data collection over decades from patients and their families has generated various types of data, including clinical, imaging, molecular, and multi-omics data. Dr. Kalish highlighted the challenges of studying BWS, such as its mosaic nature and the need for large patient cohorts and emphasized the potential of FaceBase to link and augment existing data sets. She detailed a case study of a patient with multiple BWS features and explained how data from this patient contributed to several studies, including cancer screening guidelines, molecular mechanisms of overgrowth, and the development of induced pluripotent stem cell (iPSC) models. Kalish also presented preliminary findings on tongue differences between BWS subtypes and the potential for using facial imaging to further understand BWS to inform diagnosis. She concluded by emphasizing the need for FaceBase to merge and compare BWS data with controls to better understand the disease mechanisms and develop interventions.

FaceBase and Multimodal Healthcare Data Integration (Panel)

The final segment of the meeting was a panel on leveraging FaceBase in related health domains. The panelists featured Dr. Ben Yixing Xu (USC), a glaucoma specialist who is using the underlying platform used for FaceBase (DERIVA) in the EyeAI project. Yang Chai, Jennifer Kalish and Axel Visel also participated as panelists. The discussion focused on the challenges and opportunities of working with multimodal data in healthcare, particularly in ophthalmology and other domains. The speakers highlighted the importance of integrating various data types, such as imaging, electronic health records, and genetic information, to improve disease detection and diagnosis. They discussed the need for standardized data models, such as OMOP, to facilitate data harmonization across different institutions. The panel also discussed

the potential of AI and telemedicine to address healthcare disparities in underserved populations.

The panel focused on the challenges and opportunities of integrating diverse data types into platforms like FaceBase, emphasizing the importance of human expertise over relying solely on AI. Participants discussed the need for clear data standards, the evolution of ontologies, and the balance between comprehensive and minimal data collection. They highlighted the role of community input in defining data quality and integration processes, with a focus on making data usable while maintaining flexibility for future updates. The discussion also touched on the potential of generative techniques for creating reference datasets and the importance of addressing variability in clinical grading.

Next Steps

The FaceBase team identified the following next steps as a result of this meeting and from the Internal Advisors Meeting that immediately followed the Forum:

1. Strategic Data Partnerships & Large-Scale Ingests

- **EarBase launch.** Complete migration of 350 TB of temporal-bone images from four NIDCD sites and stand up the new EarBase portal on FaceBase.
- **CranioRate federation.** Publish API + schema to exchange craniosynostosis data with the CranioRate project.
- **HEAL repository onboarding.** Finalize infrastructure tweaks that let HEAL Initiative investigators deposit pain-related datasets.
- **Human genomics ingestion.** FaceBase currently holds genomics data along with imaging and other phenotype data, however, not everyone realizes that we can handle genomics data. We will explore ways to highlight FaceBase's capabilities as a 'one stop shop' for both imaging, genomics, and other data types. The scope of this work will need to be discussed by the Steering Committee.
- **Investigate crowdsourced inflows.** Explore feasibility of pilot pipelines for already-consented data (e.g., All of Us, CHOP morphometrics).

2. Clinical & EMR Integration

- **TMD pilot.** Standardize ICD-10 diagnoses, consent language, and data dictionaries, and link records to imaging. In the future, scale Dr. Sedghizadeh's cohort from 2K→10K cases and include multiple sites.

- **MyTMD & MyDocNote apps.** Help Drs. Khalifeh and Clark package patient-reported outcomes and structured notes for repository submission.

3. AI-Ready Analytics & Tool Upgrades

- **Face2Gene / GestaltMatcher bridge.** Investigate potential to support interoperability endpoints for phenotype-driven ML diagnostics.
- **Produce guidelines for synthetic & AI-generated reference datasets.** Support rare-disease benchmarking.
- **AI-assisted curation bot**. Extend the current language-model prototype to accelerate metadata completion.
- **Refresh legacy tools.** Systematically update the Human Genomics Analysis Interface and 3-D Facial Norms tools with Mary Marazita and colleagues.

4. Platform Security, Access & UX

- **CADR 2.0 + RAS login.** Finish NIST 800-53–aligned security upgrades and enable NIH RAS single-sign-on for controlled-access data.
- **Minimal-information model evolution.** Continue developing our data model to capture the most essential data points of most use to users, as driven by community feedback + SGX3 usability test results.
- **Clinician-friendly UI overhaul**: Update the UX to surface search facets and dashboards tailored to clinical users.

5. Community Engagement & Governance

- **Feature-ranking survey.** Send surveys to users allowing them to rank proposed new areas of development, to help set priorities and avoid scope-creep. Also develop other methods for gathering more input from the community.
- **Ongoing DAC policy refresh.** Conduct regular reviews of controlled access data policies to stay ahead of re-identification risks in genomic and facial-image data.
- **Living documentation & ontology program.** Provide continuous updates to contributor guidelines, plus mechanisms for expert/community review of evolving classifications.

6. Incoming Data & Model Systems

- **Avian models.** Upload CPP / Talpid bulk-seq data and explore cross-species comparative genomics resources.
- **Beckwith-Wiedemann Syndrome.** Kalish lab to submit linked clinical, imaging and multi-omics datasets; FaceBase to extend submission schema as needed.
- **EyeAI collaboration.** Work with Dr. Ben Xu to prototype ophthalmic-craniofacial data integration.

- **Philter de-identification tool.** Request documentation on this tool from Dr. Thomas Peterson and consider collaboration on a pilot on clinical notes for FaceBase.