2024 FaceBase Community Forum Summary

Executive Summary

The 2024 FaceBase Annual Community Forum, held on May 7 & 8 at the main NIH campus in Bethesda, MD, highlighted the strategic alignment between FaceBase and NIDCR data management and sharing strategies, celebrated significant achievements in data diversity, and outlined plans to facilitate the reuse of scientific data, including electronic health records (EHR). These are a culmination of our efforts to achieve data-driven solutions for everyone's health. The Forum showcased impactful case studies on cleft lip/palate and craniosynostosis, emphasized ethical data sharing practices, and explored the benefits of data reuse and collaborations.

Dr. Jennifer Webster-Cyriaque, acting Director of the NIDCR, emphasized NIDCR's data science strategy, underscoring the importance of robust data management and sharing practices to improve trust and quality in funded research results. The FaceBase team reviewed its significant accomplishments, including increased data diversity through an open community model and ambitious plans for broadening contributions of data in related research domains and support for Artificial Intelligence (AI) and Machine Learning (ML) capabilities.

Guest speakers were organized around two disease models, cleft lip/palate and craniosynostosis, and featured case studies demonstrating FaceBase's role in foundational research, data science, and clinical studies. The "Best Practices in Data Stewardship" session delved into gold-standard methodologies for data sharing and management as outlined by NIH policy and supported by FaceBase. We also learned about important ethical considerations in data sharing.

Collaboration examples like EnamelBase showcased the successful integration of specialized research resources through strategic partnerships. Dr. Kenneth Yamada of NIDCR provided a sterling example of FaceBase's utility for Intramural research as he presented salivary gland data recently deposited into FaceBase. Additionally, the Forum highlighted FaceBase's potential as a training resource for future scientists, including a comprehensive roadmap for future training programs.

This year's event included two important panels that were central to the Forum's discussions. "Bridging Foundational, Translational, Clinical, and Public Health Research" highlighted the critical role of data reuse in various research areas, emphasizing time and cost savings, validation, and the integration of clinical data, including EHR. The "Expanding FaceBase to Accommodate Research Needs for Integrated EHR, Imaging and -Omics Data" panel also addressed the potential integration of EHR through pilot projects to expand FaceBase's translational research impact. Discussions focused on improving data-linking algorithms, standardizing consent language, and developing best practices for patient consent.

Introduction and FaceBase Update

The Community Forum kicked off with a presentation by Webster-Cyriaque that highlighted NIH and NIDCR's data science strategies and FaceBase's role in implementing strategies, enumerating the following goals:

- ○ Improve capabilities to implement the NIH Data Management and Sharing Policy (DMSP)
- ○ Develop programs to enhance human-derived data for research
- ○ Provide new opportunities in software, computational methods, and AI
- ○ Support for a federated biomedical research data infrastructure
- ○ Strengthen a broad community in data science

Dr. Rob Schuler (University of Southern California), team leader of FaceBase, reviewed FaceBase's accomplishments over the past five years, including moving from a closed hub and spoke model to an open community model emphasizing self-curation. At the time, FaceBase was not certain if the community would embrace open contributions, but submissions and releases have been steadily increasing over time and have provided high-quality curated datasets. FaceBase has also increased the breadth, depth and diversity of its data by recruiting dental and oral research (not just strictly focusing on craniofacial research) and adding new model organisms (e.g., chick). FaceBase also expanded support for research across the translational spectrum and successful pilot projects brought in translational, clinical data, including the COHRA caries research data from Dr. Mary Marazita (University of Pittsburgh). The FaceBase infrastructure is also now 100% on cloud services, enabling even further growth and expansion of capabilities. We have also enhanced our data model to make our datasets AI- and ML-ready, with scalable bulk data import and export pipelines, and integration with additional community ontologies.

Looking forward, FaceBase announced plans (provided a renewal is approved) to expand collaborations for translational data science, provide a curation pipeline to streamline clinical data submission, and build new capabilities around AI/ML (including an ML pipeline for automated phenotype prediction and a type of FaceBase Chatbot to provide more semantic search capabilities). FaceBase also demonstrated the capability to support other NIH institutes in relevant domains with a proof-of-concept demonstration for hosting audiogram data.

Impactful Uses of FaceBase Data in Research

This year's Forum held sessions with a focus on two disease models - cleft lip/palate and craniosynostosis - to demonstrate research across the translational spectrum (specifically, foundational research, data science, clinical studies/public health initiatives). Invited speakers indicated how FaceBase has and will serve as a rich resource across many different domains. These sessions were moderated by Dr. Zubaida Saifudeen (NIH/NIDCR).

The Cleft Lip/Palate session speakers were Dr. Jeffrey Bush (UCSF), Dr. Junichi Iwata (TBD) and Devin Feigelson (Operation Smile).

The Craniosynostosis and Skull Formation session speakers were Dr. Ethylin Jabs (Mayo Clinic), Dr. Antonio Porras (University of Colorado Anschutz and Children's Hospital Colorado), and Dr. Jesse Goldstein (University of Pittsburgh).

Highlights from these talks include:
- Seeing how diverse areas of the FaceBase resource were integral to research that was presented:
  - Bush's talk included annotated images of secondary palate development in mice, from FaceBase's "Mouse Anatomy Atlas" resource, as part of his study of the effects of the Tgds gene.
- Through FaceBase's "Mouse Data Summary", Iwata used gene expression data in the palate for E12.5, E13.5, E14.5 and E15.5 in his bioinformatics analyses (along with other data from MGI, dbGaP) for experimental validation.
- Iwata made some recommendations for improving reuse of FaceBase data:
  - In-house QC checks by data scientists
  - Uniform data descriptions
  - Q&A communication tools
  - Handshake platform
- Jabs described her spoke project in FaceBase Phase 2: the Transcriptome Atlases of the Craniofacial Sutures project. This project aimed to create comprehensive gene expression atlases of 11 major craniofacial sutures at embryonic stages in wild-type mice and craniosynostosis mouse models using bulk RNA-Seq and single-cell RNA-Seq assays.
  - An example of this project's impact is a study using spatial transcriptomics to reveal the role of sensory nerves in preserving cranial suture patency through BMP/TGF-β signaling modulation, identifying 2,515 genes with enriched expression in dorsal root ganglion neurons (which reused Jabs' suture single cell RNAseq data from FaceBase.)
- Porras emphasized the challenges in balancing data sharing with patient privacy. His research was based on retrospective data that was not consented for secondary research (and therefore challenging to share outside of his institution). He was able to generate models based on the data that are still useful for data analysis without compromising patient privacy (there is more related discussion about this topic in the "Bridging Data Reuse Across the Translational Spectrum" section below.)

In another session - "Reusing FaceBase Data" - moderated by Alicia Chou (NIH/NIDCR), speakers specifically addressed how they reused FaceBase data in their research:

- Dr. Aki Ushiki (UCSF) utilized FaceBase's mouse and human genomics data for a Massively Parallel Reporter Assay (MPRA) project, exploring the genetic and phenotypic variation in musculoskeletal structures. Ushiki emphasized the value of FaceBase data for MPRA library design, highlighting its role in understanding the genetic causes of craniofacial diseases and validating GWAS variants.
- Dr. Emmanuel Aladenika (University of Iowa) analyzed FaceBase data, alongside data from various other sources, to investigate rare variants within craniofacial enhancers. This research identified deleterious variants that could contribute to the risk of non-syndromic clefts, demonstrating the utility of FaceBase data in studying transcription factor binding sites and genetic risk factors.

Making the Case for Excellent Data Stewardship, Collaborations and Training

Dr. Noffisat Oki (NIH/NIDCR) moderated the "Developing and Implementing Best Practices in Data Stewardship" session, which focused on the importance of data management practices, examples of collaborations and a vision for expanding FaceBase to train the next generation of DOC researchers and clinicians.

Best Practices in Data Stewardship and Bioethics

Alicia Chou (NIH/NIDCR) walked attendees through the importance of the updated NIH Data Management and Sharing (DMS) Policy by making good data management practices a requirement. The objectives of this policy include promoting trust in the data published by NIH-funded research and improving transparency. Chou outlined the elements of the plan and provided resources to learn more on the NIH website.

Laura Pearlman (University of Southern California), from FaceBase, went into detail about best practices in managing and sharing data, including important standards established by FAIR and TRUST protocols among others. She also related these standards to how they are all supported in the FaceBase infrastructure. Pearlman provided resources on how FaceBase can assist with writing DMS plans via individualized templates and boilerplate text (if you choose FaceBase as your repository of choice.)

The session on ethics in data sharing, conducted by Drs. Carrie Heike and Stephanie Kraft (Seattle Children's Research Institute), taught us the importance of balancing ethical considerations with the sharing of data, and how they incorporated patient, family and community perspectives into the data sharing practices of the Craniofacial microsomia: Accelerating Research and Education (CARE) study.

Collaboration between a UG3 Consortium and FaceBase

EnamelBase, a UG3 consortium tasked with creating *amelogenesis imperfecta* models, worked with FaceBase to integrate the results of their research into a resource for the community (https://www.facebase.org/resources/enamelbase/). This minisite includes a primer on *amelogenesis imperfecta*, 28 new mouse models, and 20 related protocols. Drs. Jan Hu (University of Michigan) and Derk Joester (Northwestern University) described EnamelBase, each of the EnamelBase projects, and the benefits for the community.

Bringing Intramural Data to FaceBase

FaceBase contributors have typically been funded from the Extramural Research branch of NIH/NIDCR. With the location of this year's Community Forum centered on the NIH campus, this was an excellent opportunity to engage with Intramural investigators directly to help them learn more about FaceBase as an important resource and destination for sharing their data. The FaceBase team held a tutorial specifically for Intramural researchers the day before the Forum and walked attendees through the FaceBase website, the types of data and resources available, how to access them and the process for contributing data.

Earlier this year, FaceBase started working with Dr. Kenneth Yamada, an NIH Distinguished Investigator in the Intramural branch of NIDCR, to contribute our first salivary gland dataset (which was released just after the Forum - https://doi.org/10.25550/4Y-GHEM). Yamada gave a stunning talk on this data, "RNA-seq Analysis of Murine Salivary Gland Development", about how salivary tissue recombination can modify cell fate and opportunities for data sharing in terms of salary glen development. His FaceBase dataset includes raw data with gene expression and imaging data and hundreds of differentially expressed genes; however, Yamada pointed out that his lab has only focused on a few of them. Additionally, he shared that using FaceBase for data sharing provided significant cost savings compared to his previous use of FigShare, which had cost thousands of dollars. With FaceBase, he could share his data freely using FAIR principles. There are many opportunities for bioinformatics analysis of this data and Yamada hopes the community will investigate further using the shared data on FaceBase.

Using FaceBase to Train Future Scientists

In her session at the FaceBase Forum, Dr. Christy McKinney (University of Washington) discussed and outlined a comprehensive roadmap for the significant potential of FaceBase as a training resource for future scientists from diverse backgrounds. She described the essential characteristics of a translational scientist, emphasizing roles such as process innovator, domain expert, skilled communicator, systems thinker, rigorous researcher, team player, and boundary crosser. McKinney highlighted the importance of teaching trainees to utilize existing resources like FaceBase to formulate and test hypotheses. Training formats proposed include web-based sessions tailored to specific domains, team science workshops, week-long bootcamps, and single-day workshops integrated with national meetings like the American Cleft Palate Craniofacial Association (ACPA) and the American Association for Dental, Oral, and Craniofacial Research (AADOCR).

McKinney also discussed outreach strategies to investigators, training grant PIs, and relevant NIH institutes to raise awareness of training opportunities. Programmatic outreach involves RFAs and supplements for FaceBase projects and the development of training programs using FaceBase data, with a focus on integrating other NIH-funded datasets and repositories to create a comprehensive resource for trainee projects and research.

Bridging Data Reuse Across the Translational Spectrum

In the panel titled "Bridging Foundational, Translational, Clinical, and Public Health Research – Lessons Learned and Future Perspectives," moderated by Dr. Ethylin Jabs, many speakers from earlier in the Forum discussed the critical role and challenges of data reuse in various research areas. Iwata emphasized the significant time and cost savings of reusing FaceBase datasets, particularly for new investigators and trainees. Feigelson highlighted the importance of data as a validation tool, while Porras and Dr. Jay Patel (Temple University) underscored the importance of data reuse in advancing research on rare conditions and linking EHR to identify systemic health issues.

Challenges in data quality and privacy were prominent themes. Patel discussed the substantial effort required to ensure data quality, and Porras raised concerns about privacy and the complexity of sharing identifiable data. Porras suggested that current guidelines might need

revision to balance privacy with the need for data sharing and proposed using models based on original data and publishing summary statistics when proper consent is unavailable.

From a public health perspective, Feigelson emphasized the importance of data reuse for validation and inclusion of underrepresented populations. Marazita advocated for integrating and publicly sharing phenotype data to create a comprehensive resource for hypothesis generation and application.

The panel identified several action items, including improving data-linking algorithms, standardizing consent language for data sharing, and developing best practices for patient consent. There was a call for a structured conversation within the research community to create a framework that promotes data reuse while addressing privacy concerns, potentially facilitated by NIDCR's standardization efforts. The panel concluded with a consensus on the need for continuous evolution in data sharing practices, balancing innovation with ethical considerations.

Options for EHR Integration into FaceBase

The panel "Expanding FaceBase to Accommodate Research Needs for Integrated EHR, Imaging and -Omics Data", moderated by Dr. Vidhya Venkateswaran (NIH/NIDCR), explored the potential for integrating EHR with the FaceBase data repository to enhance translational research by linking craniofacial research data with EHR. This integration aims to increase data granularity and utility.

Marazita emphasized that supplementing FaceBase data with EHR can improve genomic studies on craniofacial traits. Dr. Thankam Thyvalikakath (Indiana University) highlighted examples of linking dental and medical records for studies on root canal outcomes and Sjogren's disease, showcasing the potential for hypothesis generation in basic science. The panel stressed the importance of the bidirectional flow between basic science and clinical data.

Dr. Carl Kesselman (University of Southern California and co-PI of FaceBase) emphasized that full integration of EHR into the FaceBase repository is not realistic; however a narrowly focused use case for providing clinical elements (such as summary statistics) will be more realizable while still being useful.

Dr. Yang Chai (University of Southern California and co-PI of FaceBase) and Dr. Lu Wang (NIH/NIDCR) discussed pilot projects leveraging clinical data and EHR integration to strengthen studies, noting challenges in approvals, data quality, and legal permissions. They suggested focusing on specific research questions to ensure meaningful integration.

The panel highlighted the need for curating datasets, standardizing data, and developing training programs to facilitate mutual learning. Basic researchers in the DOC space would receive initial training in bioinformatics, while bioinformaticians would gain familiarity with DOC basic research data, fostering a two-way exchange of knowledge. They proposed starting with a focused pilot based on the COHRA data from Marazita's lab, beginning with summary statistics and expanding gradually from there.

In summary, addressing EHR in FaceBase requires careful planning, pilot projects, and solving legal, technical, and educational challenges. The approach should focus on specific use cases, existing research infrastructure, and workforce development to ensure success.

After the Community Forum adjourned, the FaceBase team met with their scientific advisors and NIDCR program staff to recap the meeting and discuss action steps.

Action Items

Data Management and Sharing (DMS) Plans

1. **Leverage DMS Plans:**
   a. Ensure FaceBase continues leveraging DMS plans to capture more users and contributors and align with the NIDCR data science strategy.
   b. Ensure that DMS costs are clearly outlined in budgets and that review panels understand the real costs of data deposition, curation, storage, and management.
2. **Increase Advertisement:**
   a. Enhance communication and advertising efforts, highlighting the role of FaceBase as an associated repository.
   b. Collaborate with NIDCR to communicate the expectations and benefits of using FaceBase for data sharing.
3. **Proactive Engagement with Journals:**
   a. Encourage journals to direct submissions to FaceBase by proactively engaging with them and emphasizing the importance of data availability.

Expanding Data Scope and User Base

1. **Broaden Data Types and Scope:**
   a. Focus on including more clinical data and ensuring balanced representation beyond craniofacial data.
   b. Push for the integration of diverse data types, including salivary data, microbiome, and metagenomics data, to demonstrate systemic health implications and engage related NIH institutes about contributing resources.
   c. Promote the fact that FaceBase accepts clinical data and accommodates diverse data types.
2. **Highlighting and Promoting Data Use:**
   a. Consider creating a "results" database for processed data to accommodate diverse user needs, providing users with p-values for meta-analysis.
   b. Highlight and promote research articles or commentaries showcasing the use of FaceBase, potentially featuring one paper each month.
   c. Track recipients of letters of support to see if they have been funded and if they have deposited their data in FaceBase.
3. **Increase Communications Efforts:**
   a. Find more ways to advertise FaceBase's existence and amplify communication efforts to increase awareness among researchers.

Enhancing FaceBase's Ecosystem and Culture

1. **Track and Incentivize Data Use:**
   a. Track and publish how often datasets are accessed and how often FaceBase DOIs are cited, providing incentives for researchers to contribute to FaceBase.

2. **Creating an Evolving Ecosystem:**
   a. Work on scaling the FaceBase ecosystem to handle an increasing number of datasets, aiming for a culture that demands and responds with adequate resources.
3. **Pilot Projects for EHR and Broader Data:**
   a. Set up a focused pilot project with Marazita's lab to explore elements of EHR integration, such as summary statistics.
   b. Consider pilot projects for data related to systemic diseases, including salivary gland, TMD, and microbiome, to attract related NIH institutes.

Publications and Community Engagement

1. **Publish Review Articles:**
   a. Write and publish a review article on FaceBase's accomplishments over the last five years, outlining future directions and summarizing progress.
   b. Aim to publish concurrently with other NIDCR Working Groups to maximize impact.

Administrative and Logistical Considerations

- Explore ways to accommodate various business models for integrating new data and providing analysis resources without resorting to a fee-for-service model.
- Ensure resources are effectively brought in to support these initiatives.