22

precision**FDA**: A Sandbox for Innovative Collaborative **Omics Advancement**

Samuel Westreich[1], Ezekiel Maier[2], Adrienne Phifer[2], Omar Serang[1], Elaine Johanson[3]

[1]DNAnexus, [2]Booz Allen Hamilton, [3]FDA Office of Digital Transformation/Office of Data, Analytics, and Research

Abstract

precisionFDA is a collaborative, high-performance computing environment that brings together data scientists, regulators, and the global scientific community to analyze multi-omics and real-world datasets and advance regulatory science to improve public health. precisionFDA provides access to data analytic and artificial intelligence (AI) research capabilities, multi-omics applications, shared workspaces for secure data-sharing, virtual workstations supporting relational databases, RStudio, SAS Studio, Jupyter notebooks, and Public Data Challenges. **It advances omics regulatory science, best practices, and research through public Challenges and evolving informatics and AI capabilities.**

Since 2017, precisionFDA hosted 41 Challenges and received over 850 submissions, thereby engaging the public and scientific community to optimize innovative statistical, bioinformatics, and AI solutions to advance regulatory science. Omics-focused Challenges include assessing the accuracy/consistency of identifying genomic variations (Olson et al., 2022), identifying/correcting accidental mislabeling of biospecimen samples (Boja et al., 2018, Yoo et al., 2021), identifying pathogens in sequencing data (Sichtig et al., 2019), and benchmarking indel calling pipelines for oncopanel sequencing (Gong et al., 2024).

precisionFDA also supports FDA AI exploration, including running an AI Challenge series. precisionFDA provides capabilities to advance FDA omics and AI research, including a sandbox to fine-tune generative AI large language models, and integration with the FDA Intelligent Decision Lab and Ecosystem (FIDLE) GovCloud platform to empower data transfer into secure FDA environments. precisionFDA Challenges advance omics regulatory science and help establish best practices/standards. Additionally, precisionFDA supports future omics research by providing a generative AI sandbox environment that can be used for integrative analysis and interpreting multi-omics data.

Introduction

precisionFDA is a cloud-based, high-performance computing platform where large datasets are hosted, managed, and analyzed in a secure environment. The precisionFDA platform launched in 2015, and has grown significantly since then across many dimensions, including users, data, challenges and app-a-thons, functionality, outreach, and impact. Additionally, precisionFDA is a FISMA/FedRAMP Moderate Authorized Cloud Service for collaborative regulatory science and submission pre-validation. In 2020, the program moved from genomics research to multi-omics production environment.

Public Challenges

precisionFDA acts as a catalyst for knowledge generation by convening community Challenges and app-a-thons—which galvanize dialogue and scientific discovery—around innovations that advance the science of precision medicine and inform regulatory science.

Informatics and AI Capabilities

precisionFDA provides a unique environment for encouraging innovation through use and development of informatics and AI capabilities. For instance, precisionFDA allows for user guided app development, as well as data and multi-omics analytics.

Materials and methods

Public Challenges - Framework

precisionFDA deploys an iterative Challenge planning process that incorporates the following components (Figure 1):

1. Challenge Ideation
2. Prepare for Launch
3. Challenge Marketing
4. Run the Challenge
5. Measure & Report Results
6. Disseminate Outputs
7. Integrate Lessons Learned

Challenges are designed to inform and advance regulatory standards related to omics, bioinformatics, real-world data, and AI, as well as push for participants to develop tangible solutions that support FDA's mission.

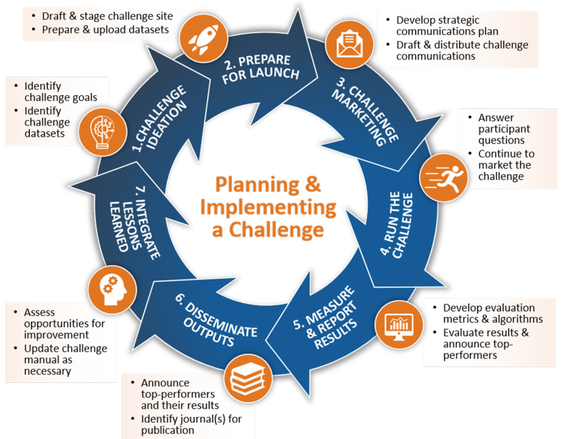

Figure 1. precisionFDA Challenge Framework

Informatics and AI Capabilities - Development

precisionFDA designs, develops, and deploys various informatics and AI capabilities, maintaining FISMA/FedRAMP Moderate requirements for all data. AI models such as OpenLLaMA can be run on precisionFDA Workstations, using either pre-trained models or running on-worker training on datasets selected by the user.

These models remain segregated on workers and do not export training data or weightings outside of precisionFDA. This allows the use of open models with restricted and secure data in a safe environment.

Results and discussion

Public Challenges – Results Highlights

precisionFDA hosted 41 crowdsourcing Challenges and app-a-thons, receiving over 850 submissions (Figure 2) throughout the years.

- 8 Challenges resulted in FDA co-authored publications
- 17 Challenges resulted in solutions/tools
- 21 Challenges advanced/informed regulatory science

Several precisionFDA Challenges have advanced omics research, including the Truth Challenge V2, the NCI-CPTAC Multi-omics Enabled Sample Mislabeling Correction Challenge, the CDRH Biothreat Challenge, and the NCTR Indel Calling from Oncopanel Sequencing Data Challenge.

Additionally, precisionFDA has hosted over 10 AI related Challenges to inform regulatory science on evolving AI capabilities.

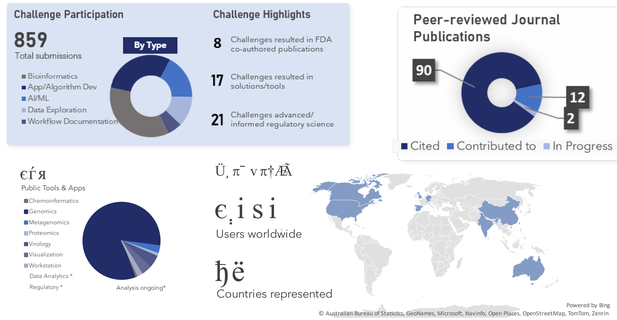

Figure 2. precisionFDA Dashboard Highlights

Informatics and AI Capabilities – Results Highlights

The platform houses a wide range of informatics and AI capabilities, including a library of applications for various use cases (Figure 3). Applications span categories of AI/ML, genomics, metagenomics, and proteomics, amongst others.

precisionFDA capabilities have been leveraged in numerous use cases such as allowing users to collaborate in a digital lab environment, perform multi-omics analysis, query and collect real-world data from trusted partners, and access the ODAR Data Analytics as a Service.

Generative AI

precisionFDA is advancing FDA AI exploration and adoption. Users are able to run their own generative AI applications on the platform in sandbox environments, choosing training data without sending data external to the platform. Additionally, generative AI large language models (LLMs) have been fined tuned on precisionFDA for FDA use cases.

FIDLE Integration

precisionFDA integrated with the FDA Intelligent Decision Lab and Ecosystem (FIDLE) GovCloud platform. This capability allows for data transfer into secure FDA environments. precisionFDA is a key component of the ODAR Data Analytics as a Service that analyzes critical supply chain information and publishes reports for the White House to predict and manage public health challenges such as infant formula shortage.

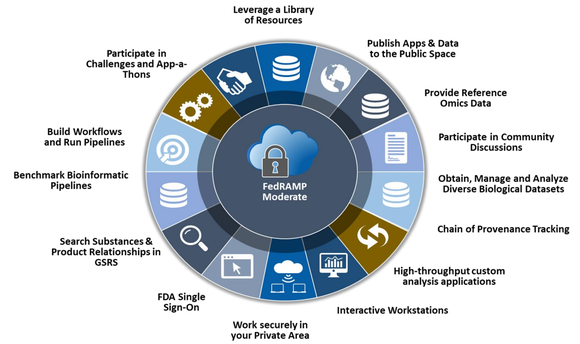

Figure 3. precisionFDA Core Capabilities

Conclusion

Public Challenges – Conclusion/Impact

Challenges and app-a-thons hosted on precisionFDA have engaged the community to develop and evaluate innovative omics, bioinformatics, and AI solutions to advance regulatory science.

A diverse network of collaborators have co-sponsored these crowdsourcing activities including FDA Centers, Government agencies (e.g., National Institutes of Health (NIH), U.S. Department of Veterans Affairs (VA)), academic institutions, patient advocacy organizations, and other public challenge hosting organizations. Further, challenges have engaged thousands of global participants and received hundreds of submissions, resulting in meaningful solutions to inform and advance regulatory science and standards, as well as define best practices.

Informatics and AI Capabilities – Conclusion/Impact

precisionFDA informatics and AI capabilities continue to be leveraged by both internal and external stakeholders to help advance regulatory science. These capabilities provide an environment for integrative analysis and interpretation of multi-omics data, supporting future omics research.