Paper AD10

CDISCDataset.com: An AI-Powered Platform for Generating synthetic CDISC-Compliant SDTM and ADaM Datasets, CRF, Protocol, TFL Shells, Specs

Ara Baghumyan, BioBrain PRO, Yerevan, Armenia

ABSTRACT

Access to realistic CDISC-compliant datasets is limited due to privacy and compliance restrictions, creating challenges for training, validation, and development of SDTM, ADaM, and TFL workflows. CDISCdataset.com addresses this gap with an AI-powered, web-based platform that generates synthetic datasets aligned with CDISC standards. Users can customize study parameters while ensuring outputs that mirror the structure of real clinical data without exposing sensitive information. By enabling secure, reproducible, and customizable dataset creation, the platform supports learning, validation, and collaboration across biometrics teams. This innovation provides the PHUSE community with a practical solution for advancing data science in clinical research while reducing dependency on real trial data.

INTRODUCTION

The adoption of CDISC standards, including SDTM and ADaM, has transformed the way clinical trial data are structured, analyzed, and submitted. While these standards ensure consistency and regulatory compliance, they also place new demands on statistical programmers, data managers, and statisticians. Teams must be able to test, validate, and refine workflows against datasets that accurately reflect the complexity of real clinical data.

However, access to such datasets is often limited. Real patient data are protected by strict privacy regulations and are typically not available for use outside of controlled project environments. This creates a gap for programmers and organizations who wish to train staff, validate processes, or test automation solutions in a compliant and reproducible manner. The lack of readily available, high-quality, CDISC-compliant datasets often slows learning, innovation, and validation activities across the industry.

CDISCdataset.com was developed to close this gap by providing a secure, web-based platform that generates synthetic datasets aligned with CDISC standards. By delivering realistic but non-identifiable data, the platform enables teams to practice, validate, and innovate without reliance on sensitive trial datasets. This paper introduces the rationale, design, and capabilities of CDISCdataset.com and demonstrates how it can support the community in advancing data science in clinical research.

CHALLENGES IN CLINICAL DATA ACCESS AND STANDARDS IMPLEMENTATION

Clinical research continues to evolve under increasing pressure to meet regulatory expectations, deliver faster insights, and ensure patient privacy. The adoption of CDISC standards such as SDTM and ADaM has provided a common framework for structuring, analyzing, and submitting data, but implementation remains resource-intensive.

One of the central challenges is the restricted availability of real clinical trial data. Due to privacy, compliance, and contractual obligations, programmers and statisticians cannot freely access patient-level data for training, testing, or validating their workflows. This creates barriers for learning CDISC standards, developing automated solutions, and performing independent validation of TFL outputs.

Equally challenging is the lack of standardized supporting materials such as study protocols, CRFs, mapping specifications, sDRG/aDRG templates, and reference SAS programs. These are often scattered, study-specific, or unavailable, which complicates training and slows the adoption of best practices. Access to reference sources, such as FDA data, is also fragmented and difficult to integrate directly into programming workflows.

Without consistent, accessible, and compliant resources, organizations risk inefficiency, longer validation timelines, and increased dependency on study-specific trial data. These gaps highlight the need for a comprehensive solution that combines synthetic datasets with the broader ecosystem of standards-based resources needed for successful biometrics programming.

OVERVIEW OF THE CDISCDATASET.COM PLATFORM

CDISCdataset.com was designed as a comprehensive platform to address the gaps faced by programmers, statisticians, and data managers in adopting CDISC standards. Rather than focusing solely on synthetic datasets, the

platform offers an integrated suite of tools that replicate the entire clinical programming ecosystem in a compliant, reproducible, and user-friendly way.

At its core, the system generates synthetic CDISC-compliant datasets across SDTM and ADaM domains, enabling training and validation without reliance on patient-level data. Beyond this, the platform provides additional resources that are critical to the biometrics workflow:

Synthetic Study Protocols and CRFs – Automatically generated mock protocols and case report forms for study simulation and training.
Mapping Specification, aDRG, and SDRG Templates – Pre-configured templates that streamline documentation and regulatory compliance.
Reference SAS Programs for TFLs – A library of reusable programs that can be selected by table type, supporting efficiency and standardization.
Synthetic EDC Raw Datasets – Realistic raw data outputs that mirror EDC structures for downstream transformations.
Integration with OpenFDA Sources – Direct access to FDA reference data to enrich programming workflows and testing.
API Integration – Secure APIs allow organizations to embed CDISCdataset.com functionality directly into their internal systems, enabling automated dataset generation, protocol simulation, and template creation within existing workflows.

By combining these components, CDISCdataset.com moves beyond dataset generation to deliver a unified platform for learning, development, validation, and innovation in clinical programming.

CONCLUSION

The CDISCdataset.com is a comprehensive platform that empowers clinical programmers and statisticians with synthetic CDISC-compliant datasets, study protocols, CRFs, mapping templates, reference SAS programs, and EDC raw data. Its integration with OpenFDA sources and secure API functionality allows seamless embedding into existing workflows, supporting reproducibility, efficiency, and compliance.

By providing realistic, non-sensitive resources, the platform reduces reliance on real trial data, accelerates training and validation, and enhances collaboration across biometrics teams. CDISCdataset.com demonstrates how a centralized, versatile solution can streamline adoption of CDISC standards, foster innovation, and improve productivity in clinical research

REFERENCES

CDISC. CDISC Standards Overview. Available at: https://www.cdisc.org/standards
FDA Open Data Portal. OpenFDA. Available at: https://open.fda.gov.

ACKNOWLEDGMENTS

The authors wish to thank the BioBrain PRO team for their contributions in developing CDISCdataset.com and supporting its deployment. Special thanks to early adopters for feedback that guided feature development.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Ara Baghumyan
Company: BioBrain PRO LLC
Address: A.Mikoyan 2/2, Yerevan, Armenia
Work Phone: +37494603608
Email:ara.baghumyan@biobrainpro.net
Website: https://www.cdiscdataset.com https://www.biobrainpro.net