PharmaSUG 2020 - Paper EP-172

TDF – Overview and Status of the Test Data Factory Project, PhUSE Standard Analyses & Code Sharing Working Group

Nancy Brucken, CSG Inc;

Peter Schaefer, VCA-Plus;

Jessica Dai, Vertex;

Cynthia Stroupe, UCB;

Dante Di Tommaso, Omeros

# Abstract

Test Data Factory is one of seven projects within [PhUSE’s Standard Analyses and Code Sharing Working Group](https://www.phusewiki.org/wiki/index.php?title=Standard_Analyses_%26_Code_Sharing). Suitable test data are an essential part of software development and testing. The objective of the TDF project is to provide CDISC-compliant data sets to empower statistical programmers and software developers. Users, primarily software developers, should be able to customize fundamental aspects of test databases.

# Introduction

Previously, the TDF project team have published two data packages based on SDTM and ADaM data sets published in a CDISC pilot. Now the TDF team have begun to implement SAS and R code to simulate clinical trial databases based on user configuration.

PhUSE is a volunteer organization that relies on community contribution to progress initiatives such as TDF. This poster and paper inform the community of TDF history, current activities and future plans. We further hope to inspire interested programmers and software developers to join our efforts, and be a part of delivering these capabilities to our industry.

# Phase 1: Completed

The TDF team has published two test data packages based on CDISC Pilot datasets:

* 31 SDTM (CDISC SDTM 3.2) datasets were updated and documented, including define.xml
* 12 ADaM (CDISC ADaM 2.1) datasets were updated and documented, including define.xml

Define.xml and Pinnacle 21 reports accompany both databases, which PhUSE have published as part of the project github repository:

* <https://github.com/phuse-org/TestDataFactory/tree/master/Updated>.

PhUSE publish a complete archive of products that PhUSE Working Groups have delivered:

* <https://www.phuse.eu/phuse-references>

While the TDF team have published these databases, we do not consider this phase truly complete until we have received candid feedback from users. We welcome all input from industry colleagues on their experiences using these updated pilot data.

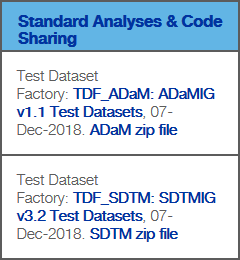


Figure 1: PhUSE publish TDF-updated CDISC databases as Working Group deliverables

# Phase 2: The TDF Roadmap

The TDF are developing a framework and syntax that allows software developers to describe basic attributes of a clinical study database, and then simulates a database according to that configuration.

## Variable Modeling

Our current focus is on modeling basic variable types:

* character variables based on NCI/CDISC controlled terminology, including sponsor extensions;
* character variables based on industry dictionaries like MedDRA and WHO Drug;
* date/time variables;
* numeric variables like laboratory results or ordinal values;

It is early days for our implementation, which is also available in the github repository referenced above.

## User INterface Collaboration

Our current user interface is based on a Trial Design Matrix macro-enabled Excel workbook. We are currently collaborating with the creators of this proprietary tool, to bring it into the public domain for our industry.

The TDF extension to this tool is a "TDF Configuration" tab that allows users to configure database details that otherwise would not appear in the Trial Design SDTM domains. For example:

* The Trial Summary (TS) tab includes required study details such as the Study Start and End Dates.
* The TDF Configuration tab provides further configuration required to simulate subject enrollment, such as the duration of an enrollment period following the study start date.

## SIMULATION considerations

Clinical study data are as unpredictable as any biological system, as anyone familiar with clinical studies has almost certainly experienced first-hand. Although a "clean and complete" clinical database is a common study milestone, our databases are typically not entirely clear or complete, despite best practices and efforts of all involved.

User requirements for test databases almost certainly include those unexpected and unpredictable data elements.

A central concept and design topic for the TDF team is "database credibility". To what extent are realistic attributes essential for a test database? Almost certainly, the more realism we attempt to simulate, the more complicated we must make both the user interface, and the platform design.

The current thinking of the TDF team asserts the following:

* Within-variable credibility is less important than across-variable and across-domain credibility
  + Within-variable credibility, for example, credible hemoglobin results for a particular study or medical population is not only difficult to achieve, but non-essential for more software development
* Reasonable across-variable and across-domain credibility is essential.
  + The interventions, events and findings associated with a simulated subject should be sufficiently consistent for common data and analysis workflows.
  + Requirements "reasonable" and "sufficient" credibility will vary broadly by project.
* Delivering a collection of credible domains is more useful that polishing individual domains.
* A solid starting point is better than nothing, and potential a better starting point than the typical industry approach, which is to de-identify or anonymize data from a prior study. Working from a solid starting point, developers can enhance a TDF database with the particular nuances that their particular project requires.

The TDF project offers many challenges.

# Conclusion

The TDF project exists to provide CDISC test datasets as an important contribution to the development and deployment of CDISC-based software solutions. Publishing the updated SDTM and ADaM packages, based on CDISC Pilot Project datasets was an enlightening first step.

An open, interactive software platform that delivers users with customized test databases remains a distant goal. Several groups such as PhUSE projects are working to deliver such a platform. We will progress according to plan, and would welcome your contributions.

# References

CDISC SDTM/ADaM Pilot Project. 2013. Accessed March 15, 2020. <https://www.cdisc.org/sdtmadam-pilot-project>

# Acknowledgments

PhUSE is a volunteer organization that relies on community contribution to progress initiatives like TDF. Team members are grateful to PhUSE for creating a dynamic industry community replete with opportunities to develop and apply industry expertise. We hope that we have inspired software developers to join our efforts, and be a part of delivering these capabilities to our industry.

# Recommended Reading

* *PhUSE Working Groups – Volunteers driving industry advances:* [*https://www.phusewiki.org/wiki/index.php?title=General*](https://www.phusewiki.org/wiki/index.php?title=General)
* *PhUSE Standard Analyses & Code Sharing Working Group – Volunteers standardizing analyses* [*https://www.phusewiki.org/wiki/index.php?title=Standard\_Analyses\_%26\_Code\_Sharing*](https://www.phusewiki.org/wiki/index.php?title=Standard_Analyses_%26_Code_Sharing)

# Contact Information

Your comments and questions are valued and encouraged. Contact the TDF project lead at:

Dante Di Tommaso

dantegd@gmail.com

<https://www.phusewiki.org/wiki/index.php?title=WG5_Project_09>

Any brand and product names are trademarks of their respective companies.