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 Inc.;
Jessica Dai, Vertex;
Cynthia Stroupe, UCB;
Dante Di Tommaso

ABSTRACT

Test Data Factory (TDF) is one of seven projects within PhUSE's Standard Analyses and Code Sharing Working Group (SACS). Suitable test data are an essential part of software development and testing. The objective of the TDF project is to provide up-to-date and CDISC-compliant test data sets to empower statistical programmers and software developers. Users, primarily software developers and quality control staff, should be able to customize fundamental aspects of test databases.

INTRODUCTION

The TDF project team previously updated data packages based on Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) databases, and including Case Report Tabulation Data Definition Specification (CRT-DDS, define.xml) documentation, originally published for a CDISC Pilot Project. 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

In 2007, CDISC published a CDISC Pilot Project submission package based on:

- SDTM Implementation Guide (IG) Version 3.1.1, and SDTM Model Version 1.1;
- ADaM Version 2.0 (no ADaM IG at that time);
- define.xml according to CRT-DDS version 1.0; and
- Operational Data Model (ODM) version 1.3.1 (draft at that time).

CDISC subsequently updated their Pilot Project in 2013, based on:

- SDTM IG Version 3.1.3, and SDTM Model Version 1.3;
- ADaM IG Version 1.0, and ADaM Model Version 2.1;
- define.xml according to CRT-DDS version 1.0.0; and
- ODM version 1.2.1.

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

- 31 SDTM datasets were updated and documented, including define.xml
  - SDTM IG Version 3.2, and SDTM Model Version 1.6.
- 12 ADaM datasets were updated and documented, including define.xml
  - ADaM IG Version 1.1, and ADaM Model Version 2.1.

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, see **Figure** 1.

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.

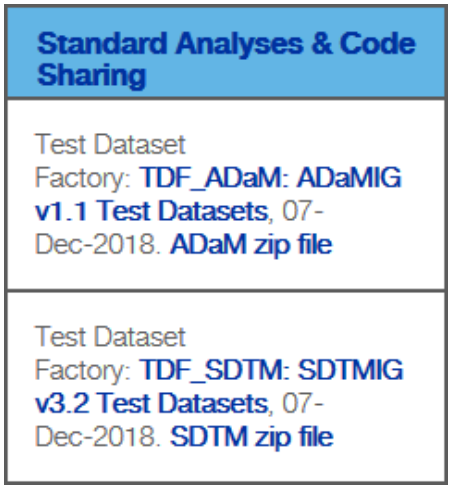

Figure 1: PhUSE publish TDF-updated CDISC databases as Working Group deliverables

PHASE 2: THE TDF ROADMAP

The next step of the TDF project is to develop a framework and syntax that allows users to describe basic attributes of a clinical study database, and then to simulate a database according to that configuration.

VARIABLE MODELING

Our current focus is on modeling basic variable types, such as:

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

Our ongoing implementation is available in the Github repository referenced above.

USER INTERFACE COLLABORATION

Our current user interface is based on the Trial Design Matrix, which is a macro-enabled Excel workbook that creates the Trial Design data sets as specified by SDTM. The sponsor that created this helpful tool is working with CDISC and the TDF project 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) domain includes study details such as the Study Start and End Dates.
- The TDF Configuration tab allows further configuration required to simulate subject enrollment, such as the duration of an enrollment period following the study start date.

Ultimately, our team wants to deliver a flexible and easy-to-use way to create custom test databases. The Phase 1 deliverables, mentioned above, provide a generic starting point, though it might require considerable effort to customize the test data packages for a particular software project. We are considering two alternatives:

- A relatively simple option would be to publish scripts (for example, SAS or R scripts) that users can use to generate their own customized test datasets. Other PhUSE projects have used this approach and one of the SACS projects has created a Github repository to publish such scripts. The downside of this approach is that the user needs to have the programming environment and have the programming experience to use the scripts. The benefit is that the scripts can be modified to perfectly meet the end user's requirements. In other words, this approach provides a flexible starting point, but also requires considerable individual effort spread across the industry.
- A more satisfying, but relatively complex solution, is to deliver a software platform to generate test databases. This solution would be available and useable to both programmers and non-programmers. We envision a cloud-based environment that would execute scripts and that would include an interface to replace the Trial Design Matrix Excel workbook mentioned above. As a proof of concept, SACS projects have created hosted R environments, accessible from standard web browsers, to host and execute R scripts. This approach allows everyone to use the scripts without the need for any programming environment, software licenses or programming expertise. A local version could be available to the community but would require an appropriate local environment within which end users could run the application. The biggest challenge of this "platform" solution, over the scripting solution, is the greater development effort, which would require more commitment from volunteers for both the initial implementation as well as ongoing maintenance.

While considering options, the TDF team will proceed with development of SAS and R scripts based on the enhanced Trial Design Matrix Excel workbook. The project team continues to seek community input and feedback to evaluate options.

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 neither entirely clear nor complete, despite best practices and efforts of all involved.

User requirements for test databases almost certainly include those unexpected and unpredictable data elements associated with clinical trials.

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 and implementation.

The current direction of the TDF team is based on a set of guiding assertions:

- 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 patient population is not only difficult to achieve, but non-essential for most software development.
- Reasonable across-variable and across-domain credibility and consistency is essential.

    - The interventions, events and findings associated with a simulated subject should be sufficiently consistent for common data and analysis workflows.
    - Definition of "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 potentially a better starting point than the typical industry approach, which is to de-identify or anonymize data from a prior study based on different assumptions and patient population. 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 design challenges, and opportunities for participants to expand their expertise by progressing solutions for each.

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 remote but appealing achievement. Other groups such as PhUSE SEND Data Factory developers do share our interests and objectives to deliver such a platform (see References). We continue to progress according to our plan and would welcome your contributions to achieve these objectives.

REFERENCES

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

CDISC Glossary. 2019. Accessed March 15, 2020. https://www.cdisc.org/standards/glossary.

SEND Data Factory PP24. 2019. FDA/PhUSE US Computational Science Symposium. Proceedings accessed March 15, 2020. https://www.lexjansen.com/css-us/2019/PP24.pdf.

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 apply and expand 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*
- *PhUSE Standard Analyses & Code Sharing Working Group – Volunteers standardizing analyses 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