

Creating CDISC Test Data Sets – A Worthwhile Concept?

Peter Schaefer, peter.schaefer@nc.rr.com

The release of the FDA guidance document about standardized study data in 2014 (see [1]) finally defines the roadmap to the mandatory use of CDISC data standards in submissions. All players - sponsors, CROs, technology providers - are reviewing their processes and new software systems are specified, implemented, and deployed. This raises the issue how to test and validate new and modified systems with realistic test data sets. After all CDISC study data are a collection of multiple files with sometimes complex dependencies and test study data that cover a wide range of options are most likely not easy to get.

The poster describes a utility that creates CDISC test datasets based on user input about the underlying study. By providing the flexibility to quickly change study parameters (sample timing, number of subjects, dosing regimen) and re-create a consistent set of CDISC domain datasets, the tool will enable faster and more comprehensive testing of processes and systems which will result in higher software quality and reliability. Some experience with an initial release focusing on study data relevant for PK analysis, specifically the domains DM, PC, and EX, is presented and discussed.

Who needs Test Data Sets?

When working on programs that handle study data, for example, for analysis and reporting, there is always the need for test data that meet certain criteria. “Some data may be used in a confirmatory way, typically to verify that a given set of input to a given function produces some expected result. Other data may be used in order to challenge the ability of the program to respond to unusual, extreme, exceptional, or unexpected input.” This nice summary from Wikipedia [2] shows why there would be demand for being able to create test data in an easy and efficient way – the requirements for test data are pretty broad and it is typically difficult to determine from real data whether certain part of a program are challenged or not.

The “CDISC DataSets” Prototype

A “CDISC DataSets” prototype has been implemented as a service (“SaaS”), i.e. no software is installed on a user’s desktop. A user would go to a website, present credentials for authentication, and then can start using the program.

The user only needs to specify certain study specific data, for example

- study design (parallel, crossover)
- dosing regimen
- and some characteristics of the subjects

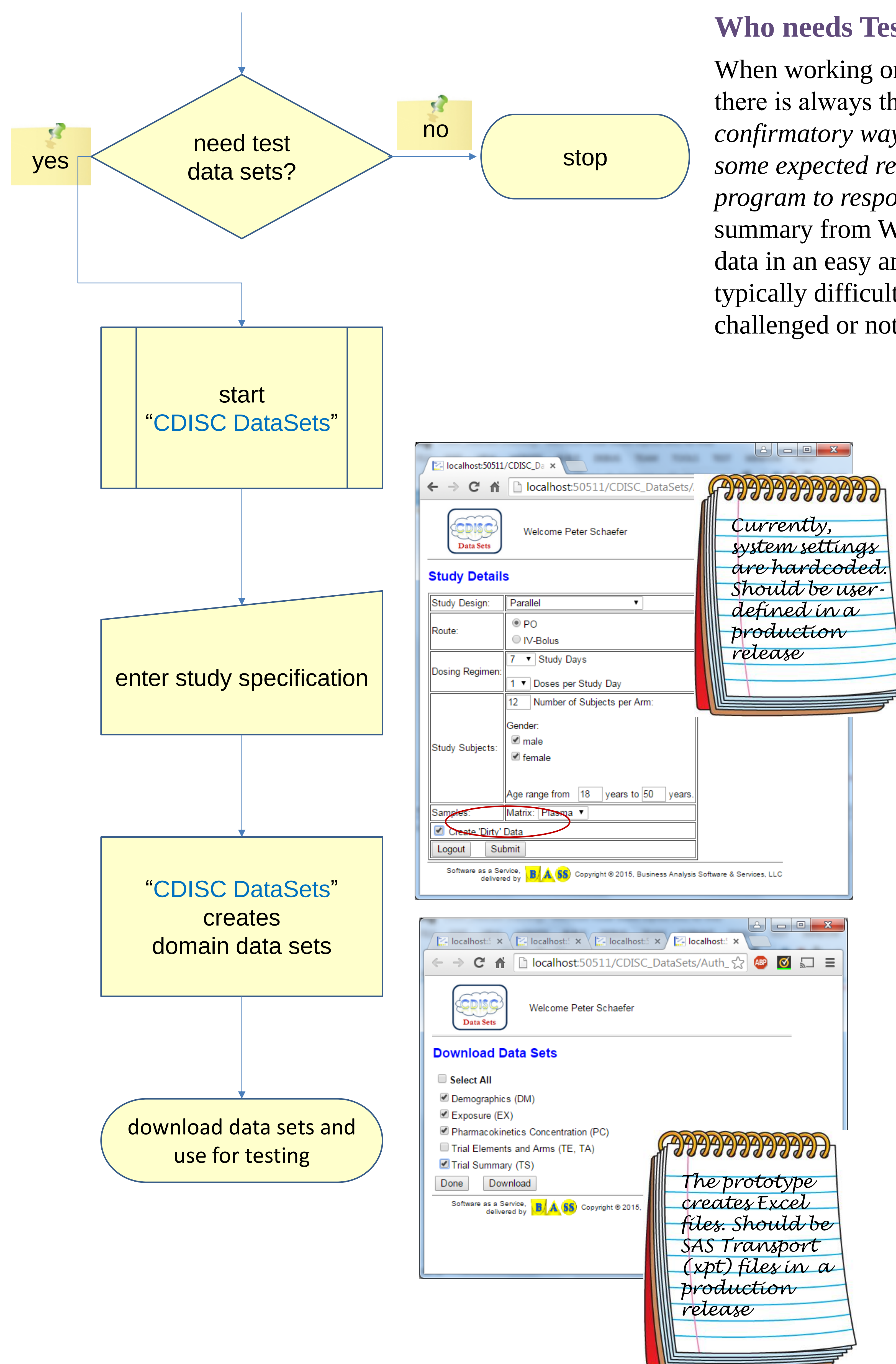
Many details of the study data are created based on system settings. This applies specifically to sponsor-defined values, like subject or site identifier, analyte or treatment names, etc. Date and time values are created relative to a system defined study start date with some randomness applied to account for the usual variability in data like sampling or dosing times. The values for subjects in the Demographics domain (like gender, and age) are created randomly within the range of the user’s settings.

The ‘Create “Dirty” Data’ Option

The program allows to create data sets that are not exactly clean. For example, data sets will include missing data, random deviation with respect to sample times for observations, and questionable or not-usable observations.

Resulting Test Data Sets

The prototype program focuses on data that are used for PK analysis, i.e. the resulting data sets can be used to test systems and processes for PK analysis and reporting. The data sets for the domains can be downloaded to the user’s desktop and then can be used for testing or validation of systems.



Current Status “CDISC DataSets”

A prototype of the concept to generate CDISC test data sets has been implemented as a website, (i.e. as Software as a Service SaaS). The prototype has demonstrated that the concept in principle works, at least for the domains that are relevant for PK analysis. With a rather small set of user input data a rather broad variety of test data sets can be generated.

However, it seems that more flexibility might be required: Currently, many values in the generated data sets are based on hardcoded system settings that in a production version a user might want to modify. User feedback will be required to determine the right balance between an overly complex user interface and the restricted flexibility of the current implementation. An improved version of the prototype will be used to gather user feedback and for some testing of the concept in real world environments.



Take a Survey

Scan the Quick Response Code and take the short, five question survey to provide feedback.

Conclusion

We believe that generating test data efficiently and effectively will be an important step in developing and deploying CDISC-based systems. This technology will allow to improve the quality of systems and to quickly identify gaps and defects. Depending on initial feedback on the prototype, a first production release of the “CDISC DataSets” program is planned for late this year and should be available for commercial use by the industry. The business concept could resemble the OpenCDISC model: A free-of-charge “community” edition of the program would be supported by a revenue generating “enterprise” edition of the tool.

References

- [1] Providing Regulatory Submissions In Electronic Format – Standardized Study Data, Guidance for Industry, <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf>
- [2] Wikipedia - http://en.wikipedia.org/wiki/Test_data