#### **PhUSE Computational Science Project Request**

1. Project Title:

|  |
| --- |
| Test Dataset Factory – Creation and provisioning of SDTM and ADaM test datasets |

1. Working Group

|  |
| --- |
| Standard Analyses and Code Sharing |

1. Affected Stakeholders(s) (i.e. statisticians, software developers, investigators, SDO):

|  |
| --- |
| * Primary stakeholders are project team of the Standard Analyses and Code Sharing working group. * Secondary stakeholders are programming and deployment teams at CROs, sponsors, and regulators that require test datasets for testing components. |

1. Project Scope:

|  |
| --- |
| Testing and using appropriate test data are an essential part of any software system or component development. Useful test data have two main objectives – they show that the system works correctly when the data are ‘good’ and they show that the system can handle data that are ‘not good’.  (Note: The meaning of ‘good’ and ‘not good’ depends on the context of the system, i.e., the test data need to be created or acquired within the system specificications to be tested.) Several PhUSE CS projects describe medical research methods, features, or processes, and some even create software components or subsystems that handle SDTM or ADaM datasets. As part of these efforts, a variety of test data are required. The typical fallback position of project teams is to use data from the CDISC pilot project and/or anonymized study data that are provided by project team members.  The proposed PhUSE Test Dataset Factory project suggests a different, more systematic approach consisting of the following activities:   1. Define an initial scope of test data, i.e., which types of datasets are required (prioritized list) and what are the desired characteristics of these datasets (for example, incomplete, missing, or wrong data) 2. Create scripts (preferably R scripts) that will create the defined test datasets with specified features through a simulation-based approach 3. Identify an infrastructure that enables users to easily use the scripts to create test datasets (preferably an extension of the Github infrastructure defined in the Standard Analyses and Code Sharing working group). 4. Define and implement a process to store and publish the test datasets and the appropriate metadata thorugh the existing channels (for example, Wiki and Github repository) 5. Collect ‘real’ study data and publish together with simulated test datasets   While the first activities are intended to apply to a systematic approach to **creating** test data with well defined features, the last one is adding a ‘real world’ perspective to the project by **collecting** existing study data. It is proposed that this should be lower priority but not ignored. |

1. Project Dependencies (i.e. Resources, staff, other projects, etc.):

|  |
| --- |
| Ideally, the project will consist of a core team of about seven active contributors (two co-leaders and at a minimum about 5 additional volunteers). The project will require input from and collaboration with other projects of the working group, specifically, the Analysis and Display White Papers project (ADW), the Repository Content and Delivery project ([RCD](http://www.phusewiki.org/wiki/index.php?title=WG5_Project_02)), and the Repository Governance and Infrastructure project ([RGI](http://www.phusewiki.org/wiki/index.php?title=WG5_Project_03)). |

1. Project Objectives and Timeline:

|  |  |
| --- | --- |
| Finalize project scope and project setup | 2 months |
| Create initial prioritized list of required test datasets | 2 months |
| Create infrastructure and work plan for creation of test datasets through R scripts | 3 months |
| Delivery and qualification of first test datasets to be used for testing central tendency scripts developed by the RCD team. | By end of 2016 |