Paper SI08

Synthesize Your SDTMs!

Realistic Simulated SDTMs When You Need Them

Matt Travell, Jazz Pharmaceuticals, Cambridge, United Kingdom
Steve Wade, Jazz Pharmaceuticals, Palo Alto, USA
Jagan Mohan Achi, Jazz Pharmaceuticals, Palo Alto, USA

ABSTRACT / INTRODUCTION

In this paper, we will demonstrate how a prototype process was developed to simulate SDTMs on an aggressive timeline first-in-human Phase I study to support the dose-escalation review process using the study protocol, EDC specs, readily available R packages, metadata libraries and prior study data as inputs leveraging a combination of open-source technologies of R, Python and Neo4j. We will show how this successfully enabled programming activities on the SAP mock shell, ADaMs, static TLFs and interactive exploratory Shiny application to be completed prior to study enrollment.

THE CHALLENGE

Jazz Pharmaceutical is a relatively small but expanding Biopharma company which is going through a rapid transformation internalizing and modernizing its data analytics capabilities plus expanding into new molecule entities. Historically, all Phase I studies were conducted externally with partner CROs resulting in no internal processes for conducting first-in-human studies with the necessary dose-escalation analysis and decision-making robust processes in place. The challenge to the newly formed Integrated Data Analytics and Statistical Programming group (IDASP) was to enable this internal dose-escalation review process by providing both timely formal SAS generated Safety/PK/PD TLFs and an interactive Shiny based application to expediate data exploratory all with-in a 2-month time timeframe. This study challenge was simultaneously presented to the Clinical Operations group meaning that data acquisition system build would need to occur in parallel to the analysis programming activities and so no viable test data would be available to support analysis programming activities prior to the data collection starting. Once data collection started there would be a 5-day window before analysis results would need to be generated.

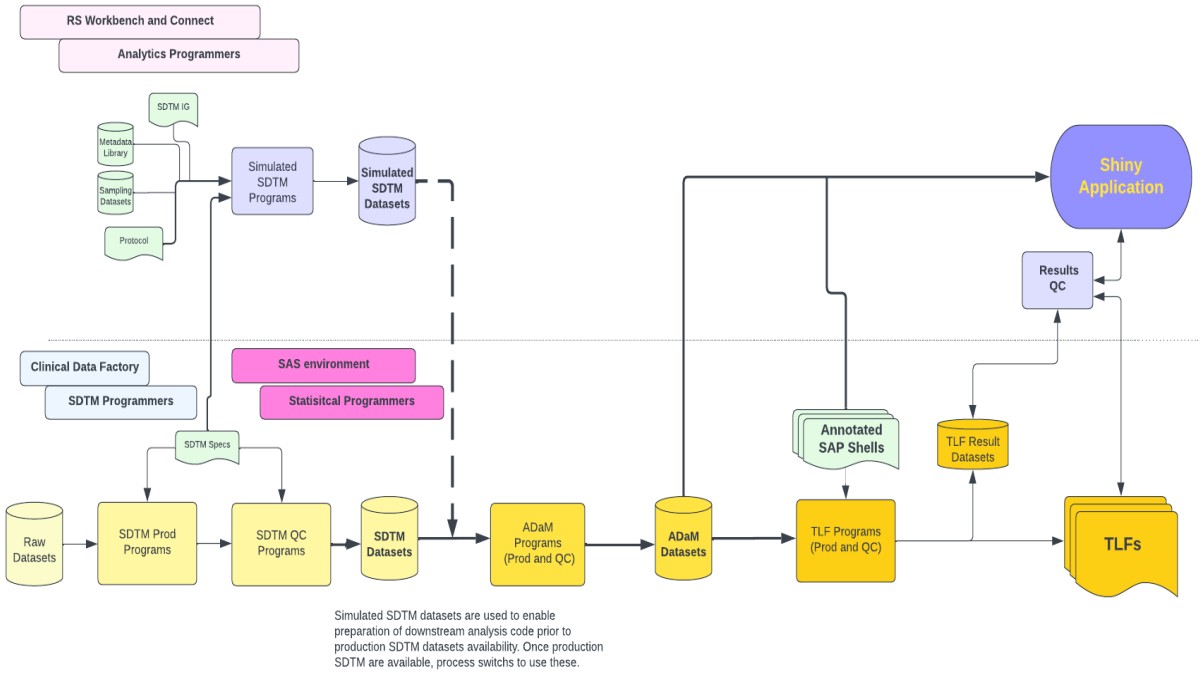
THE OPPORTUNITY

Simulation of study data had been discussed within the programming group previously as a possible option to augment and improve the test data historically available via the study set-up process. This challenge provided a specific use-case where the potential value of simulating data within the IDASP team could be evaluated and perceived benefits better understood.

THE APPROACH

The approach was to use a combination of open-source software to create a full set of SDTM datasets that are fully structurally identical to the production datasets and, more critically, closely mimic the expected different subject journeys / pathways through the study and their resulting differing realistic data profiles. The availability of this simulated SDTM set would enable all downstream analysis design, code development and testing activities of Mock TLF, ADaM, static TLF and interactive TLF Shiny activities to be completed prior to actual production SDTM data being available and support a frictionless transition of code to actual data once available. To help deliver this a full agile scrum approach was used with 2-week sprint cycles and a developer team made up from SDTM, Statistical and Analytics programmers plus Product Owner and Scrum Master roles.

HIGH LEVEL WORKFLOW



DETAILED WORKFLOW

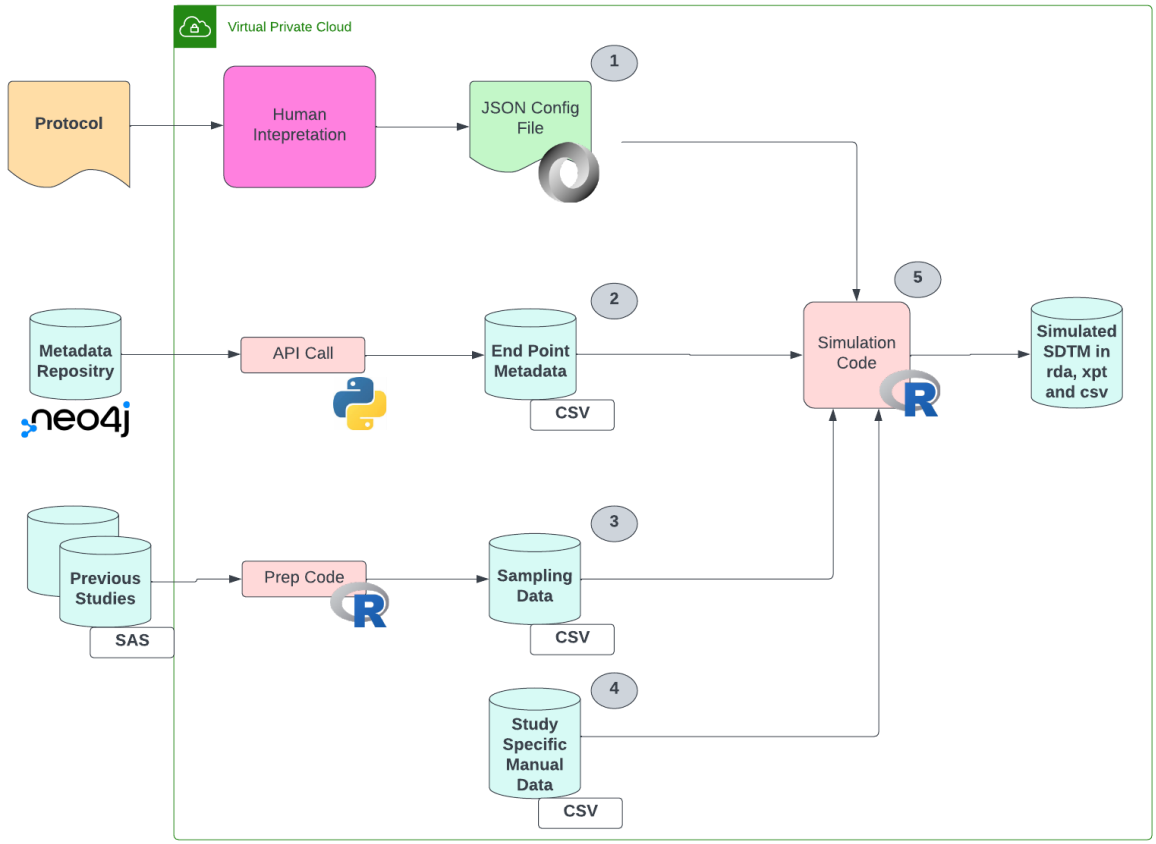


Figure 1 shows how the simulated SDTM dataset fits into and supports the regular workflow for TLF creation.

Figure 2 shows a more detailed workflow of the differing components used to create the simulated SDTMs, with the following key steps:

(1) JSON Configuration File

This is a configuration JSON text file, both human and machine readable, internally developed and designed with an intentionally simple structure to expediate the PoC process and test the use of a JSON file as controlling metadata exchange format. The file acted as the main input to drive code execution.
This included:

- Study design level information (e.g. number of subjects, treatments groups, randomization ratios)
- Population (e.g. eligibility criteria)
- End points in the study. List of end points groups (e.g. ECG) and the required end points (e.g. Holter QTcF, Lead QTcF)
- Time points in the study. Time points per end point at the collected lowest level (e.g. collection of triplicate ECGs)
- Subject pathway information per time point

(2) End Point Metadata

This is created from a central repository of end point metadata stored in a linked graph database (Neo4j AuroDB). Python code was used to connect through a bolt API to run cypher queries and pull the data into a 2-dimension CSV file formatting for processing ease.
This included:

- Target SDTM structural metadata for each end point taken from SDTM IG and study specifications.
- End point data metadata generated from legacy internal studies:
  - Distribution type
  - Estimate value
  - Variance
  - Minimum
  - Maximum
  - Categorical values and their associated probability
  - Trend of end point over time

3) Sampling Data

Sampling data for Adverse Events, Medical History and Concomitant Medications was generated by combining relevant data from legacy Phase I Healthy Volunteer studies.

4) Study Specific Manual Data

Manually created subject records to inject adverse event reaction profiles into subjects. This included adverse events records, parameters trend updates (e.g. Vitals, Labs), concomitant medications and discontinuation records.

5) Simulation Code in R

The simulation code was written in R 4.1.2 using RStudio running on a Linux server in an AWS environment using Github as the code management tool. The code flow had 3 main steps:

- Metadata set-up (packages: **dplyr, jsonlite**) where the inputs from 1+2 are combined to create controlling temporary objects.
- Simulation of results (packages: **dplyr, simstudy**) where controlling objects are used to generate actual simulated values.
- Processing into SDTM structure (packages: **dplyr, admiral, xptr**) where simulated data set is processed into final SDTM target compliant structures.

THE OUTCOME

The outcome of the PoC was that simulated SDTMs that mimicked expected study subject data profiles were successfully created within the needed time frame. This enabled the statistical programming study team to program ADaMs then work through realistic TLF Mocks (Figure 3) and interactive Shiny application design (Figure 4) discussions with key stakeholders during study set-up. This early requirement confirmation enabled code and QC testing to be completed well in advanced.

Then during study conduct, which had a 2-day window for analysis generation, this enabled a seamless switch to production SDTMs once these become available.

Figure 3 (TLF shell with simulated data)

Jazz Pharmaceuticals
Protocol JZ[redacted]

Page 1 of 1

Table 9.1.1.1
Study Disposition
Randomized Analysis Set

Total Number, n (%)	Placebo (N=22)	JZP[redacted] mg (N=20)	J[redacted] mg (N=20)	JZ[redacted] mg (N=10)	JZ[redacted] mg (N=40)	JZ[redacted] (N=20)	[redacted] (N=132)
Participant Treatment Status							
Completed	22 (100)	18 (90.0)	19 (95.0)	10 (100)	35 (87.5)	19 (95.0)	123 (93.2)
Discontinued	0	2 (10.0)	1 (5.0)	0	5 (12.5)	1 (5.0)	9 (6.8)
Adverse Event	0	0	1 (5.0)	0	2 (5.0)	0	3 (2.3)
MET QTC STOPPING CRITERIA	0	2 (10.0)	0	0	0	0	2 (1.5)
MET VITAL SIGN STOPPING CRITERIA	0	0	0	0	3 (7.5)	1 (5.0)	4 (3.0)
Participant Study Status							
Completed	21 (95.5)	17 (85.0)	18 (90.0)	10 (100)	32 (80.0)	18 (90.0)	116 (87.9)
Ongoing	0	0	0	0	0	0	0
Discontinued	1 (4.5)	3 (15.0)	2 (10.0)	0	8 (20.0)	2 (10.0)	16 (12.1)
Adverse Event	0	2 (10.0)	1 (5.0)	0	5 (12.5)	1 (5.0)	9 (6.8)
Lost to Follow-up	1 (4.5)	1 (5.0)	1 (5.0)	0	3 (7.5)	1 (5.0)	7 (5.3)

Figure 4 (exploratory interactive Shiny application with simulated data)

THE LESSONS LEARNED

The following insight was taken from the PoC:

- Perceived benefits of PoC where successfully realised and the process can be used to improve the quality of available test data. There is also the opportunity to further develop the concept across other analysis component of ADaMs, Analysis Results Data, Mock TLFs and Code.

- Need non-technical ability to digest a protocol and understand how this translates into realistic subject data profiles. Generating structurally correct simulated data is straight forward but understanding subject populations and anticipated data trends requires a skillset that may being lacking to statistical programmers.

- Technical challenges lied more in creating the configuration files and connecting the various sampling data / metadata sources across a new cloud-based platform than the simulation code itself. This was due to a good open-source package already developed available from CRAN for the simulation work (simstudy, https://CRAN.R-project.org/package=simstudy, Reference 1). While the AWS environment provides opportunity to connect multiple programming environments and databases, getting these working effectively posing a technical challenge.

- End point distribution and historical trend information was not readily available and needed to be generated.

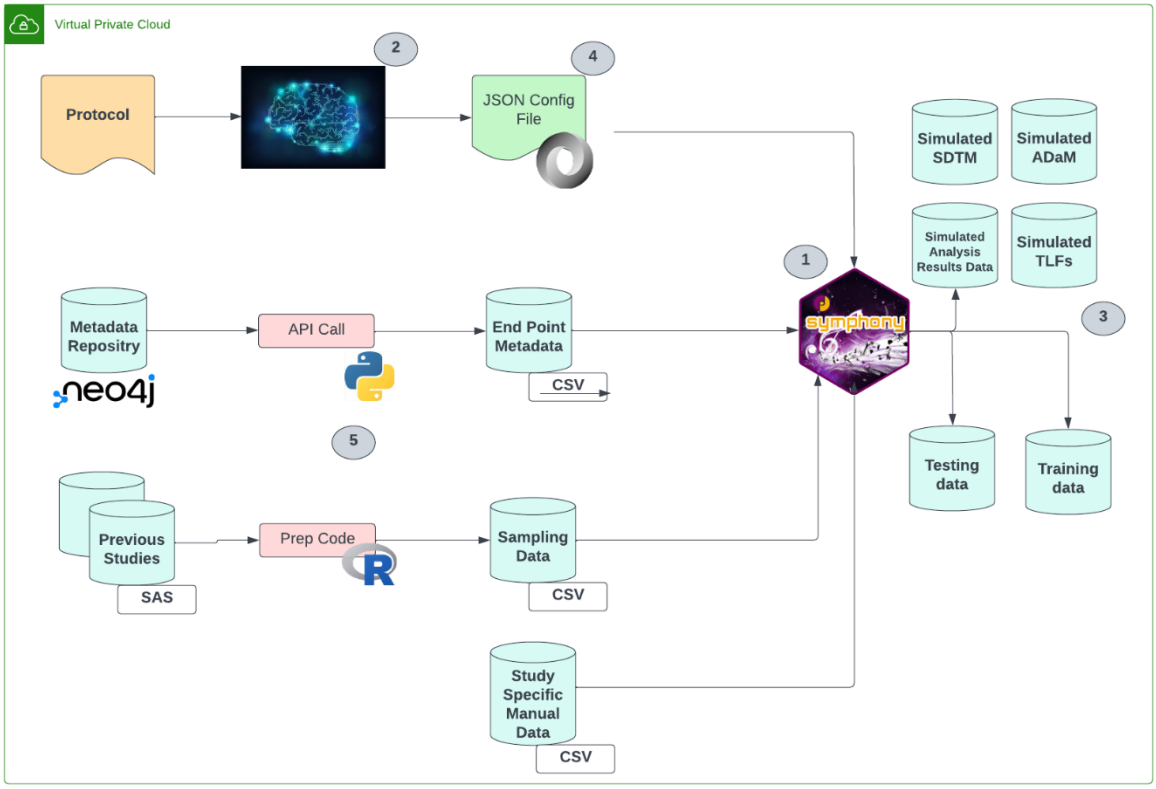
- Needed to manually inject data scenarios across data domain.

- Still need to perform the raw to SDTM programming activities. Simulating raw datasets is possible however is more problematic when none standardised data collection raw extracts are in place and/or no final EDC platform. If standard EDC extracts are in place or an earlier EDC build, then process from raw can be simulated.

- Need a good source of legacy sampling data. Would not potentially exists for new study end points.

THE NEXT STEPS

To build on the success of the PoC, the following areas have been identified as providing further value:



(1) Formulating into a user friendly and robust R package

(2) Replace the human manual step of creating the configuration JSON file with a machine learning step to automatically generate from key study documents such as the Protocol or SAP text. This would enable full automation of the process.

(3) Different use cases for simulated data

- Testing of standard functions
- Demonstration of interactive tools
- Generation of large-scale data for ML model training

(4) Investigate CDISC USDM and other industry led initiatives for aligning JSON configuration design and content expansion.

(5) Expand and formalize “knowledge management” capabilities on a new Data Science platform including Metadata Repository, legacy sampling data and connectors to publicly available data sources.

REFERENCES

(1) Goldfeld K, Wujciak-Jens J (2020). “simstudy: Illuminating research methods through data generation.” Journal of Open Source Software, 5(54), 2763. doi:10.21105/joss.02763, https://doi.org/10.21105/joss.02763.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Matt Travell
Company: Jazz Pharmaceuticals
Email: matthew.travell@jazzpharma.com