useR! Salzburg 2024

useR! Salzburg 2024

# Generate Raw Synthetic Data for Clinical Trial

Binod Jung Bogati

Senior Statistical Programmer

Numeric Mind

# Overview

- Challenges
- Use cases
- Working mechanism
- Future work

*Disclaimer: My views are my own. It doesn’t represent any organization.*

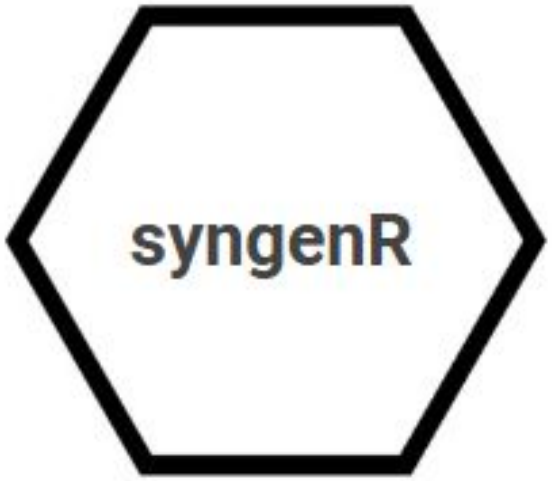
# Challenges

- Manual data entry in EDC systems.
- Difficulties in creating realistic test data scenarios.
- Issues with efficiency in raw data set generation.

# Motivation

- Testing and validation for pilot studies
- SDTM Mapping Training for R users
- Reduction of manual data creation

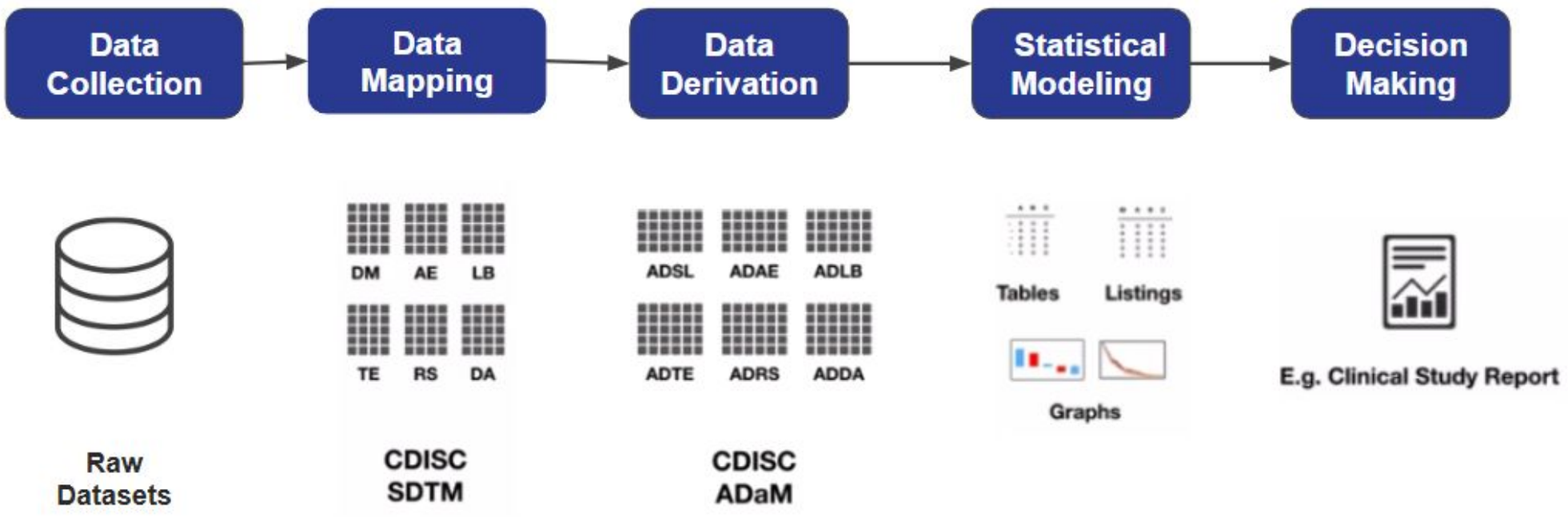
# Introduction



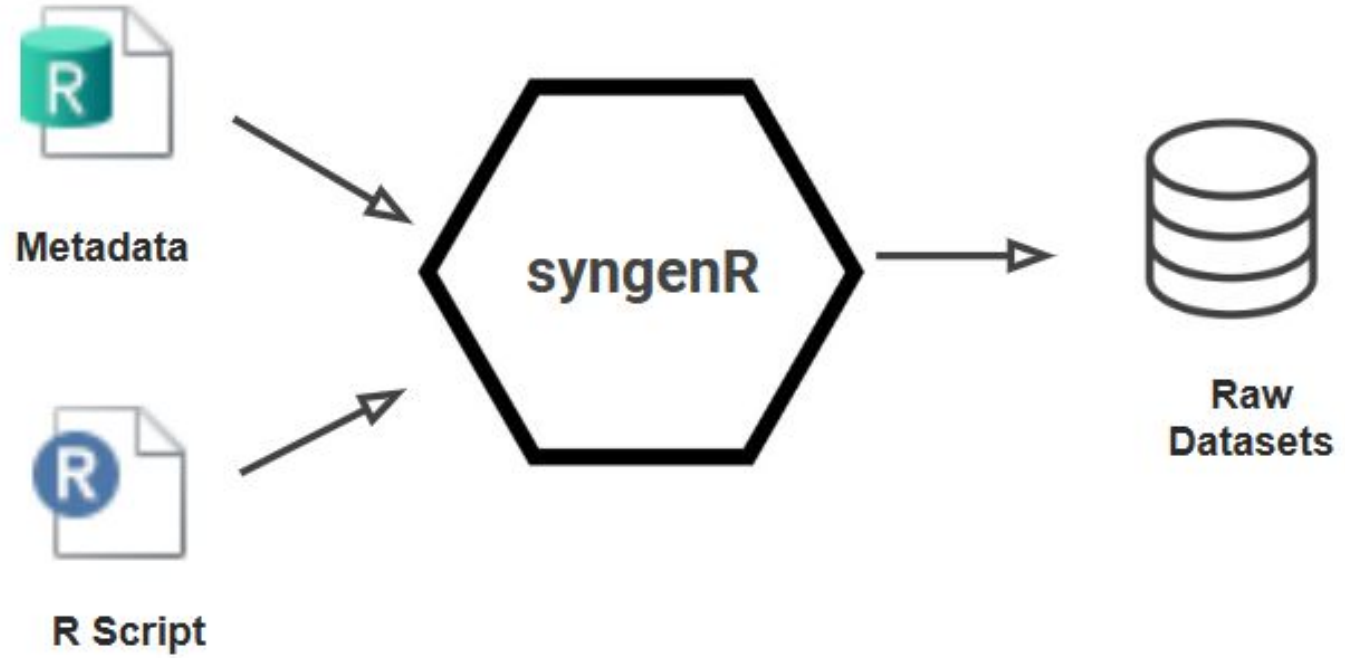

- Ready made raw synthetic data set for use
- Less manual effort to generate raw data set
- Customize data sets to reflect study needs

# Background

## Clinical Trial Programming Steps



# How syngenR works?



# Metadata: Study Information

| type | name | value |
|---|---|---|
| config | study_id | SYN-01-100 |
| config | subject_size | 20 |
| config | site_init | 1000 |
| config | subject_init | 100 |
| config | study_start_date | 10/11/2019 |
| config | study_end_date | 10/1/2023 |
| config | protocol_start_date | 10/11/2019 |

# Metadata: Visit (Schedule of Assesment)

| CRF Pages Name | Screening | Treatment | | | | | Followup | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | Week 6 | Month 3 | Month 4 6 8 10 12 |
| Informed consent | X | | | | | | | | |
| Demographics | X | | | | | | | | |
| ECG | | X | | | X | | | | |
| Vital Signs | X | X | X | X | X | X | X | | |
| Physical Examination | X | | | | | | X | X | X |
| Haematology | X | | | | | | X | | |
| Chemistry | X | | | | | | X | | |
| Coagulation | X | | | | | | X | | |
| Urinalysis | X | | | | | | X | | |
| Exposure | | X | X | X | X | X | | | |

# R Scripts: syngenR Functions

syngenR 0.1.0 Reference

## Function reference

### All functions

`bell_gen()`
Generate Bell-Shaped Numeric Data

`birth_date()`
Generate Birth Dates

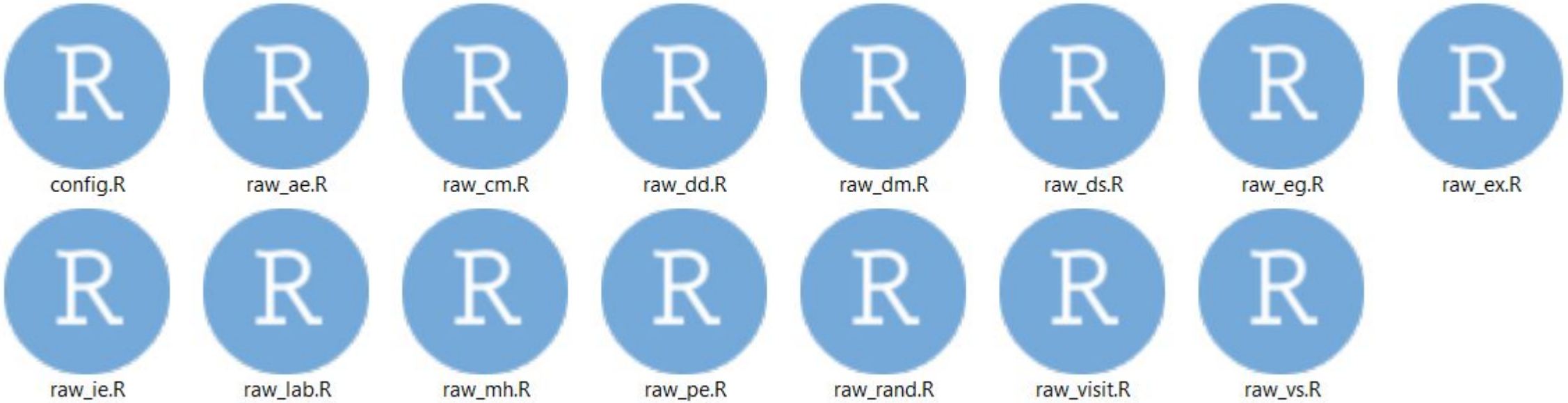
`birth_year()`
Calculate Birth Year

`cel_to_fhr()`
Convert Celsius to Fahrenheit

`date_gen()`
Generate Random Dates

```
birth_date(25, as.Date("2023-01-01"), 5)  # Generates 5 birth dates for 25-ye
#> [1] "1998-08-25" "1998-10-10" "1998-08-09" "1998-07-19" "1998-09-25"
```

# R Scripts: External



# Raw dataset Output of raw_dm (N = 20)

| subject | crfpageid | crfpagename | folder | foldername | recordid | recordposition | size | age | age_raw | birth_year | birth_month | birth_day |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1001-0001 | 1 | Demographics | SCN | SCREENING | 1 | 1 | 20 | 32 | 32 | 1988 | 12 | 20 |
| 1001-0002 | 2 | Demographics | SCN | SCREENING | 2 | 1 | 20 | 39 | 39 | 1981 | 8 | 10 |
| 1001-0003 | 3 | Demographics | SCN | SCREENING | 3 | 1 | 20 | 32 | 32 | 1988 | 5 | 12 |
| 1001-0004 | 4 | Demographics | SCN | SCREENING | 4 | 1 | 20 | 48 | 48 | 1972 | 6 | 16 |
| 1001-0005 | 5 | Demographics | SCN | SCREENING | 5 | 1 | 20 | 35 | 35 | 1985 | 9 | 1 |
| 1001-0006 | 6 | Demographics | SCN | SCREENING | 6 | 1 | 20 | 31 | 31 | 1989 | 4 | 18 |
| 1002-0001 | 7 | Demographics | SCN | SCREENING | 7 | 1 | 20 | 45 | 45 | 1975 | 4 | 6 |
| 1002-0002 | 8 | Demographics | SCN | SCREENING | 8 | 1 | 20 | 36 | 36 | 1984 | 3 | 24 |
| 1002-0003 | 9 | Demographics | SCN | SCREENING | 9 | 1 | 20 | 41 | 41 | 1979 | 2 | 11 |
| 1002-0004 | 10 | Demographics | SCN | SCREENING | 10 | 1 | 20 | 45 | 45 | 1975 | 10 | 15 |
| 1003-0001 | 11 | Demographics | SCN | SCREENING | 11 | 1 | 20 | 41 | 41 | 1979 | 3 | 8 |

# Table Output using raw_dm dataset

Table 14.1.2
Summary of Demographics
Safety Population

| Variables | Category | Statistics | Treatment A (N=15) | Placebo (N= 5) | Overall (N=20) |
|---|---|---|---|---|---|
| Age (y) | | n | 15 | 5 | 20 |
| | | Mean (SD) | 39.4 ( 7.30) | 38.4 (10.26) | 39.1 ( 7.85) |
| | | Median | 40.0 | 37.0 | 40.0 |
| | | Min - Max | 25.0 - 49.0 | 24.0 - 52.0 | 24.0 - 52.0 |
| Sex | Female | n (%) | 3 ( 20.0%) | 2 ( 40.0%) | 5 ( 25.0%) |
| | Male | n (%) | 12 ( 80.0%) | 3 ( 60.0%) | 15 ( 75.0%) |
| Race | American Indian or Alaska Native | n (%) | 2 ( 13.3%) | 2 ( 40.0%) | 4 ( 20.0%) |
| | Asian | n (%) | 3 ( 20.0%) | 1 ( 20.0%) | 4 ( 20.0%) |
| | Black or African American | n (%) | 2 ( 13.3%) | 0 ( 0.0%) | 2 ( 10.0%) |
| | Native Hawaiian or Other Pacific Islander | n (%) | 3 ( 20.0%) | 0 ( 0.0%) | 3 ( 15.0%) |
| | White | n (%) | 5 ( 33.3%) | 2 ( 40.0%) | 7 ( 35.0%) |
| Ethnicity | Not Hispanic or Latino | n (%) | 15 (100.0%) | 5 (100.0%) | 20 (100.0%) |
| | Unknown | n (%) | 0 | 0 | 0 |

DATASETS: ADSL

PROGRAM NAME: t-dm.R

# Ongoing and Future Work:

- Still in Early Phase of Development
- Support dataset for complex scenarios like Crossover Study
- Add more data quality checks in raw dataset which are visualization friendly
- Expand Raw dataset for Questionnaire and other custom datasets
- Possible collaboration with other Pharmaverse tools.

# Thank you!

Link: https://github.com/bjungbogati/syngenR

Contact: bjungbogati@gmail.com