



SEND Sanitizer: Generation of Synthetic SEND Data from Existing Studies

Susan Butler^{1,2} and Kevin Snyder¹

1. Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) 2. Oak Ridge Institute for Science and Education (ORISE)

Introduction

SEND datasets are expensive and valuable outputs used to characterize toxicity profiles of new drug candidates. However, due to their proprietary nature, the availability of SEND data for toxicity modeling and large-scale cross-study analyses is limited. A possible solution to this problem is generating anonymized synthetic datasets based on the trends of real data using the application of Monte-Carlo Markov Chain (MCMC) simulations. SEND Sanitizer is a R Shiny-based application designed for this purpose. A user can generate synthetic SEND format data with trends like one or more example studies submitted into it. If the example studies are of the same species and dose arm regiments, the application can generate a user-specified amount of .xpt files for the BW, DM, DS, EX, LB, MI, TA, TS, and TX domains. This will allow for future modeling capabilities of SEND toxicological datasets to improve the performance of non-clinical toxicity studies.

Methods

Utilizing the model SEND format .xpt file(s) uploaded by the user, the R code first ensures model data has the same:

- Animal Species
- Study Type
- Dosing Arm Regiment
- SEND format Version

If all of the above is the same, a new SEND format dataset is generated from the examples with the same number of animals in each dose group, same days of data collection, and same in all of laboratory endpoints in the study provided. Data is grouped by dose to retain dose dependent findings. A trial summary file (TS) is generated with species and study type used without identifying information.

For numerical endpoints such as body weight (BW) or laboratory values (LB), MCMC is leveraged to correlate findings between tests in an animal. By sampling a fitted distribution, correlated values such as urine volume and specific gravity’s trends will be retained in simulated animals.

For categorical values such as histopathology (MI/MA), the distribution frequency of findings and severity is sampled for each endpoint for each animal. This leads to the possibility that correlated tissue findings are lost but retains the frequency of dose dependent findings.

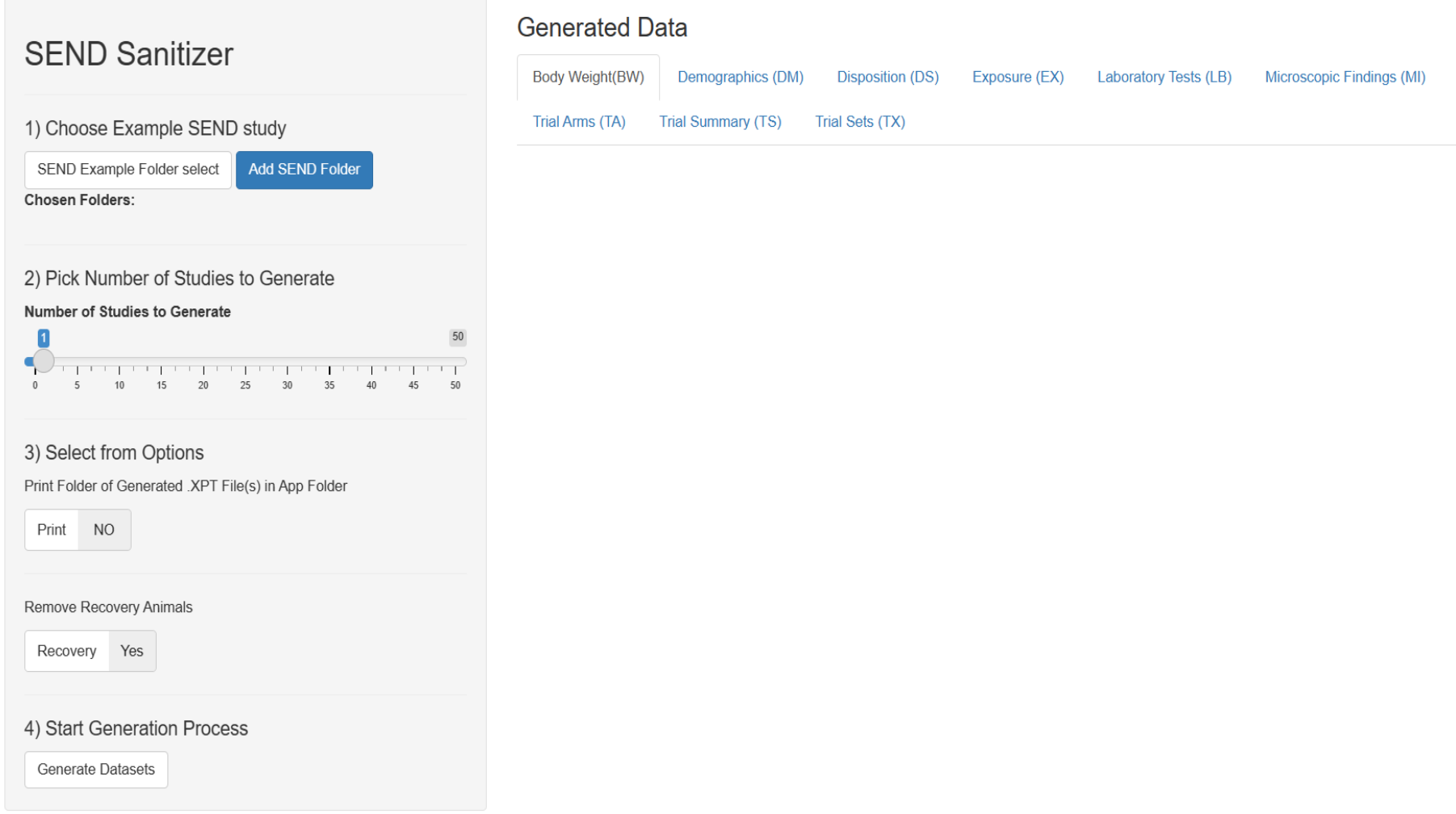


Figure 1 : R shiny interface to generate new fake studies of .xpt files of SEND format data. First, select folders of SEND .xpt files and select “Add SEND folder” for 1 or more studies. Then select number of studies to be generated and preferences for those studies in steps 2 and 3. Synthetic data will be generated after pressing the button in step 4 and displayed on the right when finished (Fig 3).

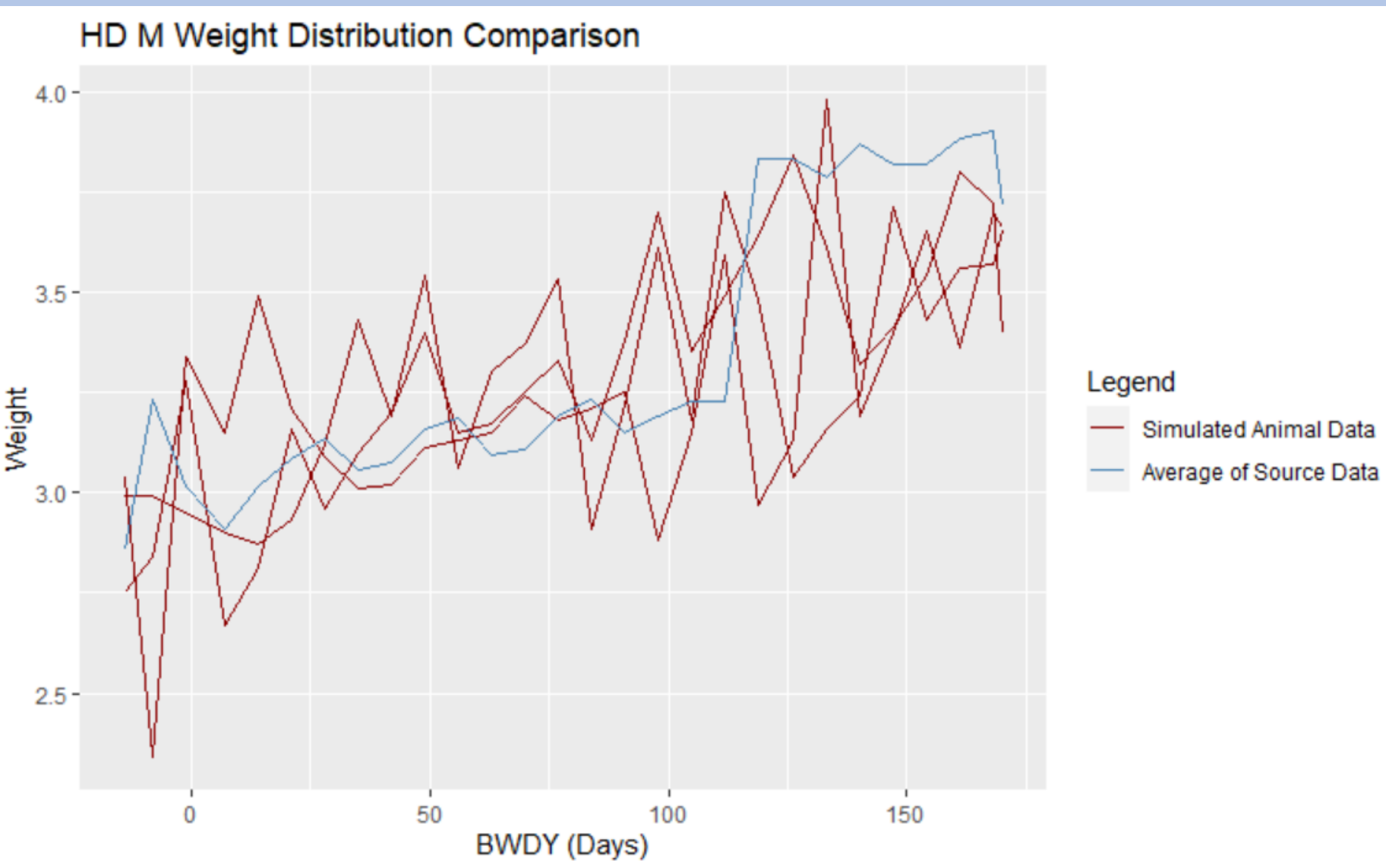


Figure 2 : Weight (kg) of 3 high dose monkey males generated by the MCMC model compared to the average high dose monkey weights used to generate the model. Simulated Monkeys follow similar weight gain trend over time to the studies used to generate them.

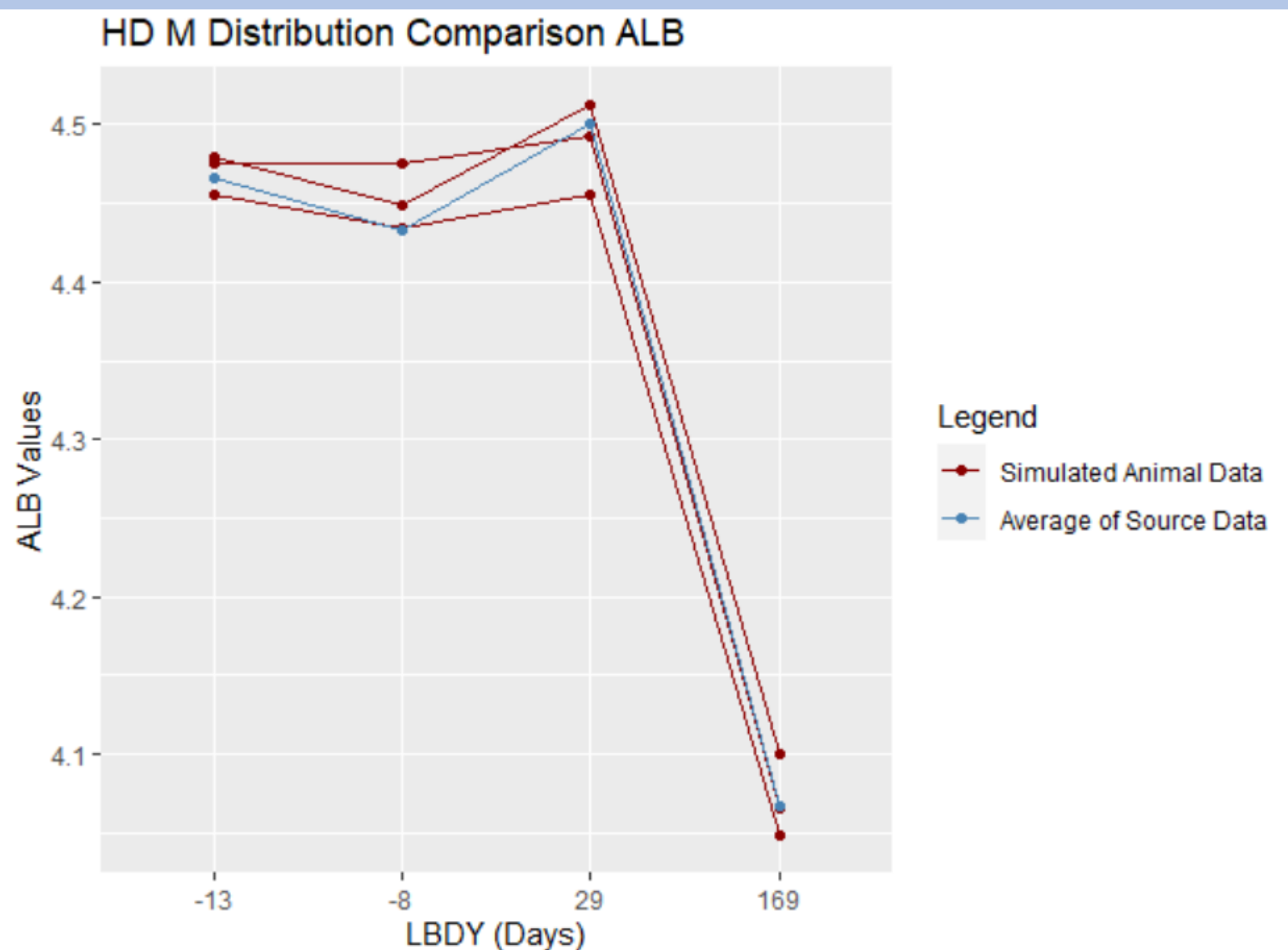
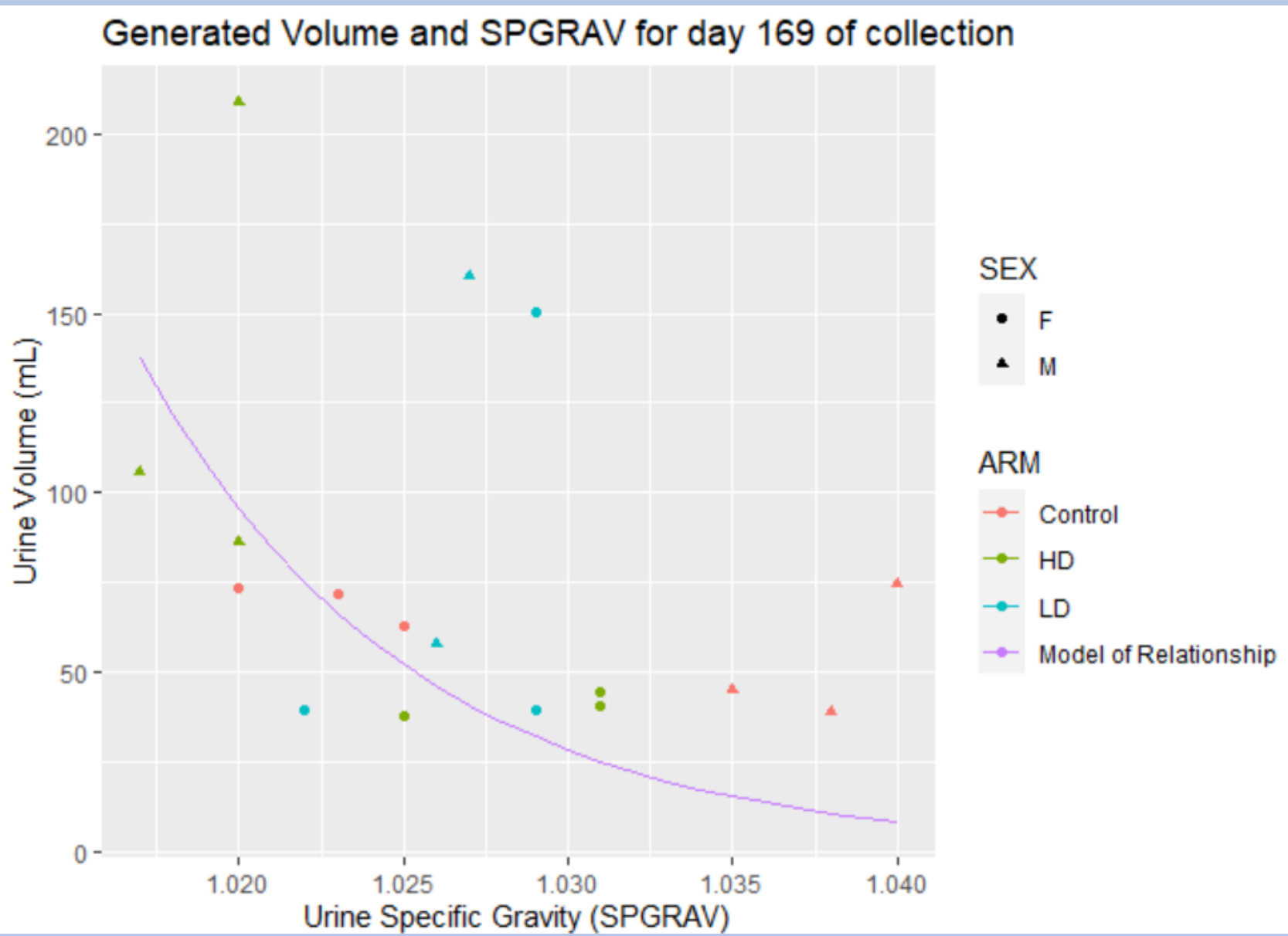


Figure 3 : ALB of 3 high dose monkey males generated by the MCMC model compared to the average high dose monkey weights used to generate the model. Simulated Monkeys follow similar ALB loss trend over time to the studies used to generate them.

Figure 4 : Negative Correlation between Urine Specific Gravity and Volume in simulated monkeys LB Data is retained. Each point on the graph is a synthetically generated male or female animal of any dose group. In real data, an exponential decay type of negative relationship exists in these points. The MCMC creates a loose version of that relationship depicted as the purple line. Further tuning of the model will improve the correlations likeness to real life.



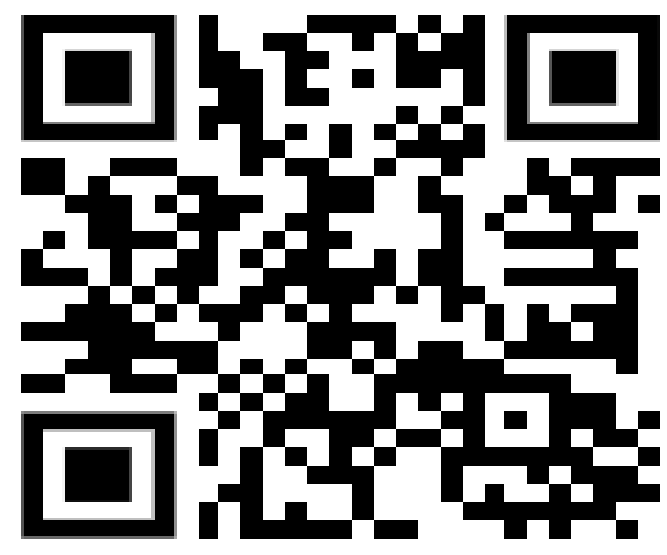
Conclusion

R Shiny can be used to generate synthetic SEND datasets from existing data. Ongoing efforts to leverage the full potential of SEND data could benefit from using datasets that have been sanitized of their identifying features for analytical use.

Future work

- Incorporation of MCMC or similar model to correlate histopathological findings inside generated animals
- Testing of synthetic data relative to original datasets used to generate it in toxicological analysis to validate if synthetic data retains toxicological significance

Code available on
GitHub!



Source Code