PharmaSUG 2026 - Paper AS-359

Beyond Imitation: Selecting Synthetic Data with Purpose and Precision

Sundaresh Sankaran, Pritesh Desai, and Sherrine Eid, SAS Institute

ABSTRACT

Synthetic data has applications for various areas of life sciences such as clinical trials, data sharing, real-world data, and data imbalances. While synthesization approaches have been variously studied as machine learning, simulation or anonymization problems, they tend to rest on assumptions involving original data distribution, which may not be sufficient or completely known. Through an illustrative example involving SAS Data Maker and SAS/OR (Operations Research) on SAS Viya, we demonstrate a method which views synthetic data as a "selection" problem.

In this method, several synthetic observations serve as contenders for selection in a desired dataset and are "selected" based on specified target objectives and rules defining constraints. This method involves augmenting conventional machine learning-based generation algorithms with optimization techniques that solve user-specified objective functions under constraints. We find two benefits from this approach - one, it guards against stochastically generated synthetic data breaking logical rules governing real data, and two, it helps tackle "unknown knowns", where only aggregations and summaries are known but not individual observations. This is especially the case when addressing rare and orphan diseases and when data access is restricted due to privacy policies and regulations.

Awareness of this method gives stakeholders access to an additional technique for synthetic data generation in scenarios involving insufficient relevant data volumes for an application.

INTRODUCTION

Enterprises face challenges and restrictions in data access, data sharing and use of real (original) data. This is due to privacy concerns, organizational policies, sufficiency and imbalance problems. For Life Sciences organisations, data is crucial for innovation but also happens to be one of the most tightly regulated and difficult resources to make available. Organisations have an obligation to guard sensitive patient data which is collected during clinical trials and operate under strict frameworks such as HIPAA. Systemic reasons, the time taken for trials to complete and imperfect enforcement of standards also aggravate the problem by contributing poor quality and incomplete data.

Synthetic data mitigates these challenges by learning from trends and patterns present in original data and ensuring they reflect in synthesized observations. This helps protect privacy by shielding real patient data from exposure, which carries legal and reputational risk.

Synthetic data generation needs to be viewed both as a technique and a process or problem to solve for. When viewed simply as a technique, synthetic data generation acquires a narrow perspective because every technique carries with it assumptions around methods, data and applicability.

For example, let's consider the simplest form of synthetic data generation – simulation of multiple patient attributes by sampling based on an assumed distribution such as normal, log-normal, Poisson etc. This may sound good in theory but runs into practical issues of generating observations that may not be observable in the real world. Thus, you might encounter synthetically generated patients with gender specified as ***female***, but who carry characteristics such as ***testosterone*** levels in reference ranges expected for ***males***, and very rare for female patients.

At the other end of the spectrum, you have increasingly sophisticated methods such as Generative Adversarial Networks, Private Bayesian Networks, and Variational Autoencoders. The challenge with such methods is that these are computationally intensive and not very intuitive for all programmers, requiring higher level of statistical knowledge and skills to tune properly. They do a better job of capturing inter-column correlations through mechanisms that seek to generate records conditional upon discrete columns (i.e. categories). This helps get more accurate results but, being a stochastic method, tend to generate observations that one may not encounter in real terms.

A COMMON CHALLENGE

While there are many approaches, a common point of frustration you encounter is as follows.

---

*Your chances of success depend heavily upon using good quality data that conforms to **your** desired data distribution.*

***But most times,** your original data or distributional assumptions (which constitute the input into the synthetic data algorithm) is usually **not** the same as that of the study you are considering.*

Examples of the many possible areas of difference you might encounter are:

1. Study explores an **innovation** in a therapeutic area that has hitherto not been considered for a given demographic.
2. Study tackles a **rare disease** for which very few previous observations exist
3. At early stages, a study's design objectives may still be **fungible** with possibility of change
4. On the data side, your original data may have been **sourced from real-world data or observational studies which do not focus on the study** or therapeutic area in question

You need to add a tool to your synthetic data generation process that offers capabilities beyond those offered by generation algorithms alone. This tool, which is best represented as a program or approach, uses optimisation techniques to align stochastic and variable output with a more deterministic state that represents the desired outcomes and objectives of your study. You also gain an added advantage through access to a transparent method that is more easily interpretable and understandable, thus improving possibility of favourable reviews by oversight teams such as a Protocol Review Board.

AN OPERATIONS RESEARCH (OR)-DRIVEN APPROACH

When we change our perspective towards the problem, we can design different approaches towards solving our data challenges.

The challenge here can be described as a case of "unknown knowns", where we have visibility into only what a dataset needs to look like at an aggregate level, but do not have the actual data observations. For example, we may know that a dataset needs to have patients with treatment adherence of minimum 80% for the information to be usable, but don't yet have individual patient data.

Another challenge, mentioned in the introduction earlier, is to ensure individual data observations make logical sense. For example, we need to weed out observations showing a patient as Male but also showing indicators of pregnancy. Or apply a logical threshold to exclude observations showing levels of blood sugar so high that they cannot be realistically observed in a living person. Depending on study objectives (for example, if the study happens to deal with a diabetes treatment), we may also want to ensure that observations for healthy patients with diabetes indicator well within reference ranges are not included (thus rendering them non-candidates for a treatment).

In this paper, we invite you to view synthetic data generation as a "selection" problem.

This implies that after you generate a dataset of synthetic data observations through established methods, you follow up with an Operations Research (OR) exercise which solves for some defined objective functions along with constraints. The data identified from these objective functions and constraints may differ from the original training data used for synthetization.

Each synthetic data record, therefore, is a candidate for "selection" under the OR problem and is assigned a binary indicator (1 indicating selection and 0 indicating non-selection).

A (VERY QUICK) OPERATIONS RESEARCH PRIMER

Operations Research (OR) is a well-established body of methods; therefore, we do not pontificate on the subject but provide a simple definition. OR involves the use of mathematical and statistical techniques to solve an equation that models a desired state of a system. It helps you answer the question: "If I desire a specified outcome, what is the extent of change I can make to the factors that determine this outcome?"

Now, let us take the above question and apply it to a clinical study project you are planning (or currently support). You might like to state your desired objective as a dataset of somewhere in the region of 1000 (synthetic data points representing) patients, with certain prior demographic conditions and existing indicators. You have arrived at this number (in reality, the number of factors is much more) based on previous knowledge of retention rates through several patient visits and want a sufficiently high number to account for those who drop out during the study. At the same time, a very high number may increase the costs of conducting the study and you do not want, say, 10000 patients because you know you are never going to ever conduct such a huge study. The number of patients you wish to solve becomes your *objective function,* while the limits or boundaries you impose on the same are called *constraints.*

Within an OR framework, you therefore frame an objective function that states:

```
Obj. Function:  Maximise (number of patients)

Subject to

Constraint 1: Study Cost should not be more than XXXXXX

Constraint 2: At least NNN patients satisfying the following criteria:- (give list of
demographic criteria here)

…

Constraint N: Your Nth constraint
```

You then apply appropriate techniques (called solvers) that arrive at the correct selection of data points (i.e. patient-level observations) that satisfy the above criteria. Based on the conditions imposed, it is likely that you may not be able to satisfy all constraints, or you may sometimes find that results show you that the solution is infeasible, i.e. it is not possible to solve for the given objective function because it's too strict or data points are limited.

SYNTHETIC DATA AND OPERATIONS RESEARCH IN TANDEM

This brings us to our design. Consider a component which we call the synthetic data generator, a model that keeps churning out synthetic data that has been trained on an original dataset (or a distribution or even in some cases a specification coded into a program). The resultant synthetic data can be considered "candidate" synthetic data but not taken into production until it is run through an optimisation component driven by an operations research program, which applies a solver to select observations that help satisfy the objective function and are thus deemed "approved" from their candidate structure.

For purposes of this paper, we use SAS Data Maker as the software component for synthetic data generation (for further verification, we also explored another algorithm available in SAS Viya called PROC SMOTE, SMOTE expanding to Synthetic Minority Oversampling Technique). For our optimisation component, we use SAS/OR on SAS Viya.

AN ILLUSTRATIVE EXAMPLE

We use an example dataset from the CDISC Pilot study which contains a comprehensive package of Study Data Tabulation Model (SDTM), Analysis Data Model (ADaM) and metadata required for a submission-style package. To maintain simplicity, length and focus of this paper, we offer one example based on the Demographics (DM) domain sampled from the SDTM data model. This approach can be extended across other datasets and other constraint types using the right mix of domain knowledge and SAS/OR skills.

Using the sample dataset provided as part of the CDISC study, we used SAS Data Maker, a focussed synthetic data generation offering from SAS to generate a large volume of synthetic data points, to serve as candidates. We can afford to do so, as Data Maker makes use of algorithms like Private Bayesian networks and Synthetic Minority Oversampling Technique (SMOTE) to train a generator model (called a generator) and use this generator many times in the future without a need to get to the original datasets. This mirrors real-life situations, where access to original data is usually tightly restricted. As a rule of thumb, we generated synthetic data records of 1.5x the volume that is required by the downstream application. This factor can be varied based on use case and selection criteria. The idea is to generate more records than necessary keeping in mind that some records may get excluded because of the optimisation procedure.

Having generated candidate variables, we then proceeded with our optimisation program. Three illustrative programs (with possibility of more getting added), representing a simple scenario and two extensions, are available at the following GitHub repository: https://github.com/SundareshSankaran/as359-synthetic-data-optimisation/.

Here is a brief description of the three programs. As this GitHub repository is a growing project, we shall continue to add more examples to the same in a gradual manner.

Name	Description
Example1.sas	Simple example which balances all RACES in the DM dataset to be ~500 observations (changeable) each
Example2.sas	Example with RACE balanced, and to enforce an average Age of 70 years (assuming the study tackles an older participation range) across all race groups

Example3.sas	Example with RACE balanced, and to use a dataset with a distribution of collection prior to study record date (DMDY in the DM domain) set at 7 days. Usually guided or suggested by study protocol criteria, this can be viewed as an efficiency exercise to help identify any poor data flow (**always relative to study design**) and may point to delayed subject registration, late demographic reconciliation or slow data capture.

How do you arrive at these objective functions? You profile your synthetic data and identify gaps between the synthetic data available vs. a desired state. For example, it is possible your study involves a new drug that needs to be tested across multiple race groups. However, previous history of RACE distributions indicate that you have poor representation (under-representation) of non-White races, as also reflected in your synthetic data.

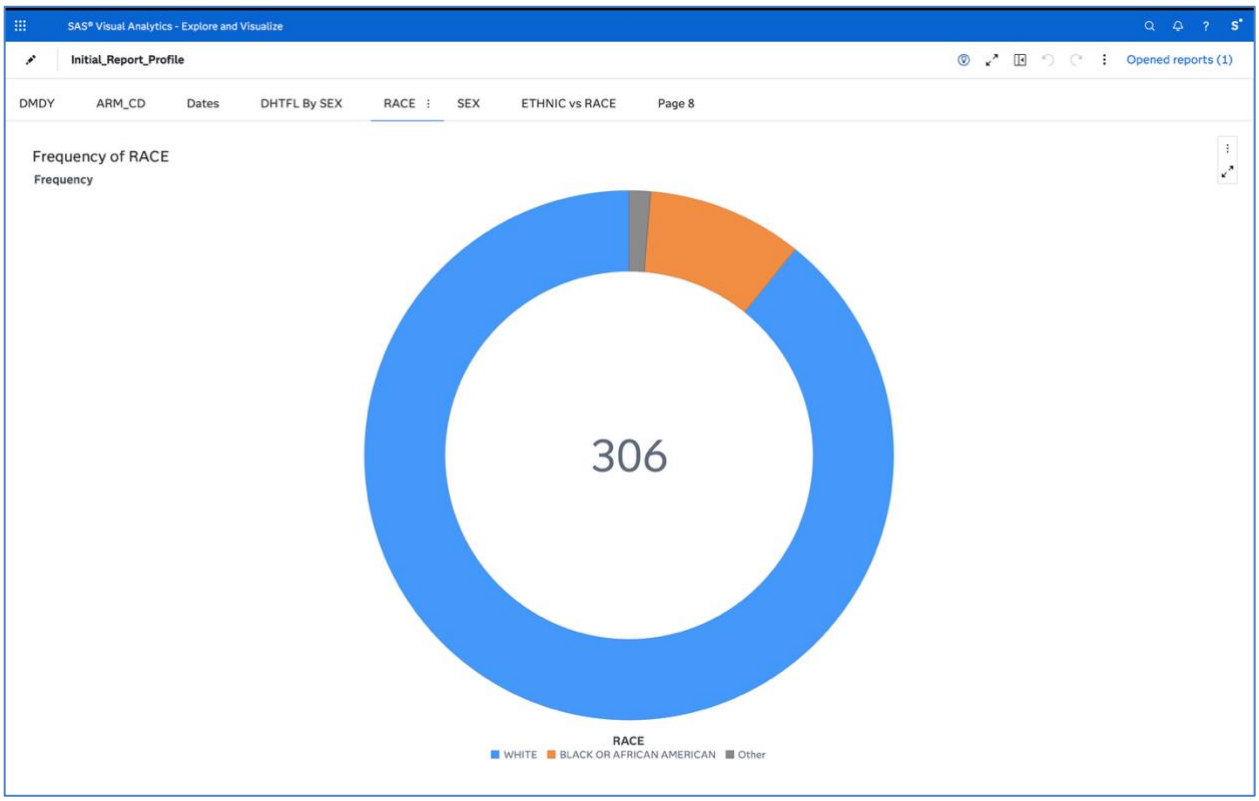

Figure 1: Original Profile of RACE on CDISC Pilot Study sample dataset (SDTM - DM domain)

Behaviour and indicators across races are not homogeneous, and you should not take the risk of proceeding with your study only based on a Race=White -heavy distribution. Therefore, you start with an objective function of establishing a (more or less) equal representation across races (Example1.sas).

In addition to the above, there might be other considerations, such as having to orient your study towards senior populations (with a higher average age than that represented by synthetic data). Yet other considerations could be operational in nature, such as a need to create test data that represents a good data flow and data capture mechanisms (as represented by DMDY, which is the study day difference between the demographics capture and a study reference date.

Using a snippet from Example1.sas (full code available in GitHub repository), we state our objective functions and constraints as follows.

```
/*************************************************************************
Define a macro variable which holds the required data observations for each
race group.
*************************************************************************/;

        %let val1agg = 500;

        /* Prints the macro variable to the log */
        %put &val1agg;

        proc optmodel;
        /* Read data into parameters */
            set<str> RACE;
            str USUBJID{RACE};

/*************************************************************************
Read in the data for the sub-group into the optimisation procedure
*************************************************************************/;
            read data DDE.DM_&i. into RACE=[USUBJID];

/*************************************************************************
Create a decision variable to indicate whether this record is selected or not
selected in the final table
*************************************************************************/;
            var x{RACE} binary;

/*************************************************************************
Objective: maximise the number of data observations per RACE group
*************************************************************************/;
            max TotalValue = sum{i in RACE} x[i];


/*************************************************************************
Constraints: Imposes a minimum and maximum condition for the number of records
which are defined within a margin of error (2% in this case).
*************************************************************************/;
        /* Margin of Errors are defined with constraints so .98 and 1.02 for example
means a 5 percent margin of error in either direction */
            con VAL1Limit: .98*&val1agg. <= sum{i in RACE} x[i] <= 1.02*&val1agg.;

            solve;

/*************************************************************************
Creates the dataset for the sub-group in the SAS library of choice.
*************************************************************************/;
        /* Output the selected observation IDs and the binary to a dataset named for the
sub-group being processed */
            create data DDE.opt_&i. (replace=yes rename = (i=USUBJID)) from [i] x;
        quit;
```

Program 1. A snippet from a simple Illustrative Program (Example1.sas) showing the optimisation procedure

SOME POINTS AND TIPS

We deliberately use a simple example to illustrate that the structure of the optimisation routine is under focus, more than the specific condition itself. Points to note are:

1. **Use macro variables to hold desired state parameters**. You'll find them easier to populate. You can also use control tables which hold these values and populate the macro variables for each run based on the condition. The following is only an example and does not make any assumptions regarding real values (which are study-specific)

Macro_Var_Name	**RACE**	**Value**
Nbr_observations_desired	WHITE	500
Nbr_observations_desired	BLACK	500
Nbr_observations_desired	ASIAN	500
Nbr_observations_desired	AMERICAN INDIAN OR ALASKAN NATIVE	500

2. **Frame your objective function as a maximisation problem.** Your objective function should facilitate "selection" of results. This selection can be optimal only if you choose to frame an objective function in such a way that all possible observations have an equal chance of getting selected in the final solution. For this purpose, frame an objective function that maximises all current observations in the dataset (i.e. select "all" observations first) and then go ahead and apply the constraints which exclude some observations from the rest.
3. **More constraints tend to hurt selection chances.** More constraints apply more restrictive boundaries and conditions on the dataset and reduce the chances of synthetic data observations getting finally selected. You need to strike a balance between ensuring that your constraints reflect relationships in your real-world data as much as possible, but at the same time, do not hinder "expansionary" approaches that might be necessary in some other cases.
4. **Focus on providing simple interfaces.** Optimisation (provided by the SAS/OR product) is a complex subject. Defining constraints in a programmatic manner may not be suitable for all skill levels. Please do explore additional wrapper mechanisms such as SAS Job Execution and SAS Studio Custom Steps that make the experience more convenient.

CONCLUSION

Through this paper, we offer an approach and an example which frames the final selection of synthetic data from a pool of candidate records as an optimisation problem and provide simple illustrative examples so that statistical programmers of all skill levels can easily get started and experiment with this approach. There are several degrees of sophistication that can be introduced to this approach. For example, you can experiment with multiple synthetic data generation algorithms ranging from simple algorithms such as random generation, to distance-based techniques such as SMOTE, and then other machine learning techniques such as Private Bayes and Generative Adversarial Networks (GANs) and explore whether the optimisation selection remains agnostic or is driven by the choice of algorithm. Further, you can also explore more complex constraints such as observations falling between a given range of dates etc. which are accomplished through min-max constraints.

An outline of the benefits from this given approach:

1. You do not need to necessarily have the exact same type of real data to get started. Even data of the same structure with similar trends and patterns is enough to make a start. This increases your time to value in being able to harness synthetic data.

2. You have access to a second level of post-processing to improve synthetic data results as originally provided by an algorithm.
3. You now have the freedom of being able to get into the area of "unknown knowns" where you have an idea of the type of data you desired but are not constrained by the fact that previous data is insufficient or may not exist. This is beneficial when considering new markets and innovation.

To recap, we list some situations where this approach tends to be most applicable:

- Where the target study differs in characteristics and objectives compared to available data from past studies.
- Where data quality issues are persistent and seep through to synthetic data and there is a need to ensure that they are weeded out.
- Where there is a need to define multiple scenarios (each containing their own objective functions and constraints) and a wish to avoid having to generate data for each scenario separately.

Also, as a pragmatic observation, let us note the limitations. This is not a magic pill. You still require some amount of original data (or knowledge of real-life data distributions and interdependencies) to achieve useful results. We suggest that you start small and with simple scenarios (much like the illustrative examples provided) and then move on to further testing with more complex criteria, some of which, for example, conditional averages, require more skill in formalisation.

Do note that such an approach may not be necessary for all synthetic data generation situations, such as when your original data already reflects the desired state to a good degree, and therefore you should be wary of falling into the trap of overengineering.

The main point, however, is that you should not feel constrained by the capabilities offered by one set of tools alone. Synthetic data generation offers huge potential for improving clinical trials and the process of generating synthetic data can be further enhanced by borrowing from related areas of optimisation. We hope you are successful in these experiments.

REFERENCES

- GitHub Repository containing examples; Sankaran, Sundaresh; Available at https://github.com/SundareshSankaran/as359-synthetic-data-optimisation/.
- Sankaran, Sundaresh & Eid, Sherrine, "Synthetic Data Generation: A Process Paradigm", PHUSE US Connect 2025, https://www.lexjansen.com/phuse-us/2025/as/PAP_AS18.pdf
- Correlation-Preserving Conditional Tabular Generative Adversarial Network Models, SAS Documentation, https://go.documentation.sas.com/doc/en/pgmsascdc/default/casactml/casactml_generativeadversarialnet_details20.htm
- SAS Operations Research (SAS/OR) Overview, SAS Institute, https://www.sas.com/en_us/software/or.html
- Chawla, et al., 2002, SMOTE: Synthetic Minority Oversampling Technique, https://www.jair.org/index.php/jair/article/view/10302

ACKNOWLEDGMENTS

The authors also thank our colleague Vasanth Ramdas for his assistance and collaboration in exploring this approach in other industries, highlighting the value of learning across disciplines.

RECOMMENDED READING

- *The SAS Linear Programming Solver: Getting Started (though it refers to SAS version 9.2, still relevant in terms of providing foundational knowledge): https://support.sas.com/documentation/cdl/en/ormpug/59679/HTML/default/viewer.htm#lpsolver_sect2.htm*

CONTACT INFORMATION

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

Sundaresh Sankaran
SAS Institute
Phone: +1 919 400 3266
Email: Sundaresh.sankaran@sas.com
https://www.linkedin.com/in/sundareshsankaran/

Pritesh Desai
SAS Institute
Phone: +1 919 400 3266
Email: Pritesh.Desai@sas.com
https://www.linkedin.com/in/desaipritesh/

Sherrine Eid
SAS Institute
Phone: +1 919 400 3266
Email: Sherrine.Eid@sas.com
https://www.linkedin.com/in/sherrineeid/