Mixed-Integer Optimization for Virtual Population Selection in QSP Models

InSysBio
QSP Company

Ivan Borisov[1], Evgeny Metelkin[1]

[1] InSysBio CY

INTRODUCTION

The goal of generating a Virtual Population (VPop) is to support drug development by **predicting the variability in patient responses observed in clinical trials**. Several VPop generation methods were proposed [1], [2]. Nevertheless, their practical application in real-world QSP projects is often limited by project-specific challenges, including:

- **Endpoints specificity**. Clinical efficacy endpoints often involve time-to-event outcomes, or survival data, which are not directly represented by the mechanistic model states and therefore require additional mapping or transformation.
- **Limited availability of individual patient data**. Clinical results are frequently reported as aggregate statistics (e.g., response rates, survival probabilities), while individual-level data may be limited or unavailable.
- **Computational complexity**. The problem becomes increasingly demanding as the number of endpoints and therapies grows.

OBJECTIVE

The objective of this study is to **propose an efficient approach for generating virtual populations (VPops) that match statistical measures** typically reported in clinical trials.

METHOD

The VPop generation process follows the two-step framework described in [2]:

1. Generation of a plausible population

The researcher defines ranges and distributions for model parameters and performs simulations using sampled parameter values. Each simulation, together with its corresponding parameter set, **is accepted if all model outputs fall within predefined biologically plausible ranges**; otherwise, it is rejected. The goal of this step is to **generate a large and diverse set of plausible patients**.

2. Selection of a VPop from the plausible population

Although the plausible population satisfies biological constraints, it does not necessarily reproduce clinical trial outcomes. The goal of this step is therefore to select **a subset of patients that matches reported clinical endpoints**.

This subset selection problem is formulated as a **mixed-integer programming (MIP) problem**. Binary variables $x_i \in \{0,1\}$ indicate whether a plausible patient is included in the VPop, subject to a constraint on the desired VPop size. The objective function minimizes the mismatch between simulated and experimental data across multiple clinical endpoints. Since individual patient data are often unavailable, the method **focuses on endpoints reported as cohort-level statistics**, such as means, std, quantiles, and survival data.

Below are examples of how objective function terms are constructed:

- **Endpoints reported as mean values**

Assuming that individual observations are normally distributed with known variance σ^2 the mismatch between the experimental mean μ_{exp} and simulated μ_{sim} can be quantified using a Gaussian approximation of the negative log-likelihood:

$$L_{mean} = N\frac{(\mu_{sim}-\mu_{exp})^2}{\sigma^2}, \text{ where}$$

N – is the sample size, $\mu_{sim} = \frac{1}{N}\sum_{i=1}^{N} y_i$ – the sample mean of simulated $y_1, \dots y_N$

In the MIP formulation, the VPop is defined by binary variables x_i, with $\sum x_i = N$

Substituting this into the expression above yields:

$$L_{mean} = \frac{1}{\sigma^2 N}(\sum x_i y_i - N\,\mu_{exp})^2$$

- **Endpoints reported as quantiles (and survival data)**

Experimentally reported values $v_1 < \cdots < v_m$ together with quantile levels $0 < q_1 < \cdots < q_m < 1$ define $m+1$ disjoint intervals with associated probabilities:

$$(-\infty, v_1), [v_1, v_2), \dots, [v_m, +\infty)$$

$$p_1 = q_1,\ p_i = q_i - q_{i-1},\ p_{m+1} = 1 - q_m$$

For simulated values $y_1, \dots y_N$ these intervals define counts $k_1, \dots k_{m+1}$ where each k_i is the sum of binary variables x_j corresponding to values falling into the respective interval. The total count satisfies $\sum k_i = N$. Assuming a multinomial distribution, a Gaussian approximation of the likelihood leads to the following quadratic form:

$$L_{quantile} = \left(\vec{k} - N\vec{p}\right)^T \Sigma^{-1} \left(\vec{k} - N\vec{p}\right),\ \Sigma^{-1} = \mathrm{N}\left(\mathrm{diag}(\vec{p}) - \vec{p}\vec{p}^T\right)$$

Further details on the mathematical formulation and implementation can be found in the package documentation: https://hetalang.github.io/DigiPopData.jl/dev/

RESULTS

A non-small cell lung cancer (NSCLC) model [2] was used to demonstrate the proposed approach. A **set of 1000 plausible patients was generated** using scripts provided in the supplementary materials of [2].

In the original study, individual patient data were used for VPop selection, including three endpoints for 112 patients across two treatment regimens (“drug” and “placebo”). To demonstrate applicability of the proposed method to more realistic settings, we **converted individual-level data into summary statistics**:

- **SLD_baseline**: mean and std of baseline tumor size (sum of longest diameters)
- **best_dSLD:** 25th, 50th, 75th percentiles of the best percentage change in SLD
- **PFS:** progression free survival data

The VPop selection procedure was **implemented in the VPopMIP Julia package** (https://github.com/hetalang/VPopMIP.jl). Using this implementation, the resulting VPop shows good agreement with the target summary statistics:

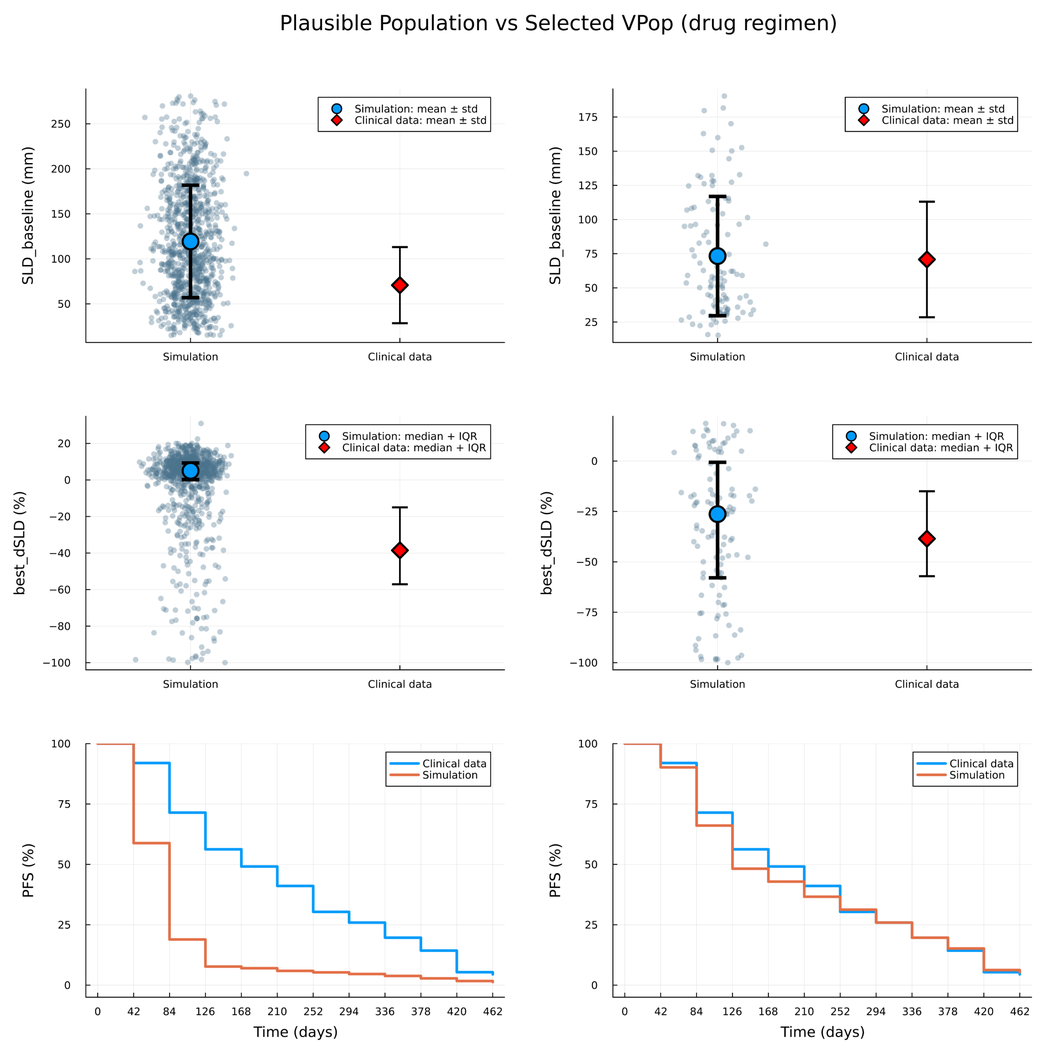

Fig 1. Comparison of summary statistics for the plausible population (left) and the selected VPop (right). The VPop contains 112 patients, matching the size of the clinical trial cohort.

CONCLUSION

1. The proposed method **enables the generation of VPops** of a desired size that **match reported summary statistics** from clinical trials. It is flexible and can accommodate **various types of statistical endpoints** by incorporating appropriate terms into the objective function.
2. The approach has two main limitations: 1) The **computational complexity** of the MIP formulation increases with the size of the plausible population, affecting performance. 2) The method relies on **sufficient diversity within the plausible population** to enable effective subset selection.
3. The resulting parameter sets can be further used for **population-level analyses**, such as inferring underlying parameter distributions or assessing parameter identifiability, etc.

REFERENCES

1. G. Kolesova, A. Stepanov, G. Lebedeva, and O. Demin, “Application of different approaches to generate virtual patient populations for the quantitative systems pharmacology model of erythropoiesis,” J Pharmacokinet Pharmacodyn, vol. 49, no. 5, pp. 511–524, Oct. 2022, doi: 10.1007/s10928-022-09814-y.
2. N. Braniff et al., “An integrated quantitative systems pharmacology virtual population approach for calibration with oncology efficacy endpoints,” CPT Pharmacom & Syst Pharma, p. psp4.13270, Nov. 2024, doi: 10.1002/psp4.13270.

CONTACTS

For more information, please visit **www.insysbio.com**
or contact the author at **borisov@insysbio.com**