Original Paper

Application of different approaches to generate virtual patient populations for the quantitative systems pharmacology model of erythropoiesis

Galina Kolesova1 • Alexander Stepanov1 • Galina Lebedeva2 • Oleg Demin1,2

Received: 26 March 2022 / Accepted: 19 June 2022 / Published online: 7 July 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

In a standard situation, a quantitative systems pharmacology model describes a “reference patient,” and the model parameters are fixed values allowing only the mean values to be described. However, the results of clinical trials include a description of variability in patients’ responses to a drug, which is typically expressed in terms of conventional statistical parameters, such as standard deviations (SDs) from mean values. Therefore, in this study, we propose and compare four different approaches: (1) Monte Carlo Markov Chain (MCMC); (2) model fitting to Monte Carlo sample; (3) population of clones; (4) stochastically bounded selection to generate virtual patient populations based on experimentally measured mean data and SDs. We applied these approaches to generate virtual patient populations in the QSP model of erythropoiesis. According to the results of our research, stochastically bounded selection showed slightly better results than the other three methods as it allowed the description of any number of patients from clinical trials and could be applied in the case of complex models with a large number of variable parameters.

Keywords Quantitative systems pharmacology • Virtual patients population • Fitting • Distribution • Mean • Standard deviation

Introduction

Drug research and development is a complex, multilayered, and long-lasting process that includes several cycles of data generation and interpretation [1]. To decrease the number of cycles and increase the confidence of the process, new techniques contributing to both data generation and interpretation have been continuously developed. Quantitative systems pharmacology (QSP) modeling [2–4] is one such technique allowing the accumulation of both publicly available and proprietary data measured in vitro, ex vivo, and in vivo. QSP modeling enables us to understand more about the pathogenesis of diseases and mechanism of action of drugs, thereby increasing the efficacy of drug development [5–7].

The conventional approach of QSP modeling includes fitting a model output to a series of mean data values. As a result, the parameters of the model represent fixed numbers enabling the description of mean data. However, the results of clinical trials include a description of variability in patients’ responses to a drug, which is typically expressed in terms of conventional statistical parameters such as standard deviations (SDs) from mean values. To allow a QSP model to reproduce the variability observed in response to a drug administration, a virtual patient (VP) population is usually generated and applied [8].

Currently, there are several approaches to generate a virtual patient population. The most commonly used is the MAPEL algorithm [7]. The main idea of the MAPEL algorithm is to introduce the so-called mechanistic axes, define different virtual patient types in the form of coordinates in these axes, and find optimal weights for each patient type to describe the empirical distribution [9]. Three other approaches presented by Rieger et al. [10] are: (1) nested simulated annealing, (2) genetic algorithm, and (3) Metropolis–Hastings. However, the virtual populations obtained using these approaches are typically large,
whereas the number of patients in real clinical trials is usually relatively small. Indeed, to obtain the best result on the basis of these algorithms, it is necessary to have rather big sample size (approximately about 1000 virtual patients). On the other hand, one could propose to select a small subset from the generated large populations. Such kind of selection cannot be done in the case of MAPEL algorithm, as the desired distribution properties are guaranteed only when the needed portions of different VP types is preserved in the population. However, small population cannot be characterized in terms of the fractions of different VP types. Three other algorithms listed above allow to perform such kind of selection. However, the exact implementation of the selecting algorithm and comparison of such a selection with the experimental results appears to be a promising direction for further research. Moreover, such methodology requires generation large VP population as initial step. Therefore, the approach will be computationally inefficient. This consideration is highly important in the case of large realistic models. Therefore, the additional requirement for the approaches developed in the present study is the ability to reproduce a population of virtual patients having the same size as that in the clinical trials.

Summarizing the arguments listed above, we can formulate the three key requirements for the approaches under development:

1. good correspondence with the clinical data;
2. ability to reproduce the population of any size (including the small ones);
3. computational efficiency enabling utilization for the cases of models with realistic large size.

We use these properties as the criteria for the comparison of the approaches proposed in the study. In particular, we formulate limitations, pros and cons of the methods in terms how well do they deal with these three tasks.

In this study, we propose and compare four different approaches to generate virtual patient populations based on experimentally measured mean data and statistics, namely, (1) Monte Carlo Markov chain (MCMC), (2) model fitting to Monte Carlo samples, (3) population of clones, and (4) stochastically bounded selection. We applied these approaches to generate virtual patient populations in the QSP model of erythropoiesis [11]. Our task was to create a sample of virtual patients of the same size as that in clinical trials and describe the experimental data. In this study we have compared these four different techniques we proposed to generate VP populations based on experimentally measured mean data and SD.

The remainder of this paper is organized as described herein. In the “Methods” section, we present a qualitative description of the erythropoiesis QSP model, the experimental data, and the algorithms of each of the four approaches. In the “Results” section, we describe the performance of the approaches with respect to the populations of different sizes. Finally, the “Discussion” section describes the advantages and disadvantages of the approaches and our future goals.

Methods

Description of model and set of experimental data

A model of normal human erythropoiesis\(^1\) was constructed to describe the cell dynamics accompanying the conversion of bone marrow hematopoietic stem cells (HSCs) to circulating erythrocytes (ERY), including the following intermediate stages: multipotent progenitor (MPP), common myeloid progenitor (CMP), megakaryocyte-erythroid progenitor (MEP), burst-forming unit-erythroid (BFU), colony-forming unit-erythroid (CFU), pre-erythroblasts (Prepro), proerythroblasts (Pro), early (Basoe) and late (Basol) basophilic erythroblasts, polychromatic erythroblasts (Poly), orthochromatic erythroblasts (Ortho), and reticulocytes in bone marrow (Ret_bm) and plasma (Ret_pl). These cellular species undergo self-renewal, differentiation, proliferation, migration from the bone marrow into circulation, and apoptosis. Essential growth factors, such as the stem cell factor (SCF), erythropoietin (EPO), and interleukin-3 (IL-3) regulate cell dynamics. The explicit binding of EPO to cell-surface receptors results in increased hemoglobin production, which in turn stimulates the degradation of hypoxia inducible factor 1 alpha (HIF) in the kidneys, thereby forming a negative feedback loop for EPO production in the kidney. Clinically relevant outputs represent levels of plasma EPO, ERY, Ret_pl, and hemoglobin as a function of ERY and RET_pl (Fig. 1).

The model was calibrated against published in vitro/in vivo data, including data for cell expansion under growth factors and cytokine exposure in vitro, flow cytometry cell counting of bone marrow aspirates, and clinical EPO administration in healthy volunteers [12]\(^2\). The model was also validated with respect to independent data sets for whole blood donation that was not used for parametrization to demonstrate the predictive power of the proposed model of erythropoiesis. For the purpose of our analysis, we used a dataset in the form of a time series describing plasma reticulocyte count in response to a single

---

1. The Human normal erythropoiesys stimulated by ESA model is available at [https://shiny.insysbio.com/ery/#/]\(^1\). Prepared by Evgeny Metelkin, Alexander Stepanov (2019)
2. Data taken from Fig. 7
dose of EPO. EPO was administered to five healthy subjects, and the reticulocyte count was measured at 16 time points. Experimental data is given in the form of a series of mean values, \( m = (m_1, \ldots, m_T) \) and a series of SDs, \( sd = (sd_1, \ldots, sd_T) \), where \( T = 16 \).

**Definition of terms used in the manuscript**

The term “control function” refers to one of the model outputs which corresponds to particular experimental data. For example, a model variable (cell or cytokine concentration) or readout (AUC of a particular variable) can be considered as a control function if the corresponding dataset describing the variable (or readout) is available. In this paper, we considered the normalized reticulocyte count as a control function. This control function represents the reticulocyte count simulated after EPO administration normalized to the baseline reticulocyte count corresponding to the model steady state before EPO administration: 

\[
\text{RET}_{\text{pl}} \text{norm} = \frac{\text{RET}_{\text{pl}}}{\text{RET}_{\text{pl}} \text{base}}
\]

**Description of the methods to generate VP population**

As a preliminary step for the VP population generation procedure, the model was fitted to the mean values obtained from literature. By doing this, we generated the so-called “reference patient”. A description of the mean experimental data with “reference VP” is presented in Fig. 2. After obtaining the “reference patient” we proceeded with the process of generating populations of VPs.

As mentioned above, the distribution of the control function should be described by the same number of VPs as specified in the clinical trial. The clinical trial dataset contains information on five healthy subjects. The reticulocyte count was measured at 16 time points. There are about 100 parameters in the model. They were either fitted or fixed at some values on the basis of available experimental data. Thirty-nine parameters of the erythropoiesis model were chosen to account for variability in the observed clinical data. These 39 parameters were chosen on the basis of empiric considerations and included several groups: (i) rate constants of the processes downstream EPO...
effect entry point, so parameters related to early progenitors before BFU do not include into variability analysis, (ii) patient characteristics (height and weight) which influence EPO pharmacokinetics, (iii) parameters related to EPO synthesis and degradation, (iv) parameters regulated EPO-dependent erythroid progenitors’ differentiation, proliferation and apoptosis. The parameters and ranges of these variations are described in Table 1. Choice of min and max values for each parameter was based on the results of large series of numerical simulations of the model. These values, therefore, reflect the ranges of the parameters allowing to avoid irregular model behavior. To implement our approaches, we used either the continuous uniform distribution for the ranges indicated in Table 1 (Approach 4) or the truncated normal distribution for the parameters with \( \mu_i, \sigma_i \), \( i = 1, \ldots, 39 \), where \( \mu_i \) is fixed on the level fitted in the model for the “reference patient”, \( \sigma_i = \frac{\mu_i}{2} \), which is bounded by zero in the left and there are no bounds in the right (Approach 1).

Therefore, our task was to create a sample of five virtual patients to describe the experimental data and compare the proposed approaches. To obtain a more precise description of the performance of the approaches, we conducted additional calculations for a relatively large population size of 207 patients. The comparison of the methodologies is presented for both the cases—one with a small population size (5 patients) and another with a relatively large population (207 patients).

The main objective of all the approaches presented in this study is to propose a methodology for generating a VP population, which describes the data control function distribution characteristics (mean and SD). Each of the approaches has specific advantages, disadvantages, and limitations. A comparison is provided in the Results section. All the approaches described in the study were implemented in R language, the source code can be accessed from the following GitHub public repository: https://github.com/insysbio/vp-erythropoiesis

![Fig. 2 Fitting of the control function (normalized reticulocyte count) to mean value in experimental data [12]. The parameter values resulting from this procedure correspond to those of the so called «reference patient»](image)

**Table 1** Physiological ranges of variable parameters

| Parameter name | Min  | Max  |
|----------------|------|------|
| k_{epo_sc_lym} | 0.02 | 0.1  |
| k_{epo_lym_pl} | 0.02 | 0.1  |
| k_{epo_sc_pl}  | 0.002| 0.01 |
| k_{epo_bm_pl}  | 0.1  | 12   |
| k_{epo_per_pl} | 0.1  | 1    |
| k_{epo_pl_bm}  | 1    | 3    |
| k_{epo_pl_per} | 0    | 0.1  |
| k_{deg_epiR}   | 1    | 40   |
| k_{int_epiR}   | 0.01 | 2    |
| k_{deg_epiR}   | 0    | 0.02 |
| k_{syn_epiR_cfu_e} | 5 | 15   |
| k_{syn_hif1a}  | 500  | 1500 |
| k_{syn_epi_mrna} | 0.2 | 2    |
| k_{base_syn_epi_pl} | 0.1 | 1    |
| k_{deg_epi_pl} | 0    | 0.5  |
| v_{max_deg_epi_pl} | 0 | 50   |
| k_{max_epi_pl} | 100  | 500  |
| k_{min_deg_epi_pl} | 50 | 120  |
| k_{BH}         | 1.5  | 2    |
| k_{max_pro_cfu_e_bm} | 2 | 7    |
| k_{EC50_pro_epi_cfu_e} | 0.01 | 0.06 |
| k_{EC50_dif_epi_cfu_e} | 0.005 | 0.05 |
| k_{Emax_dif_epi} | 1    | 10   |
| k_{max_dif_cfu_e_preproe_bm} | 2 | 6    |
| k_{base_dif_bm} | 0.05 | 0.15 |
| k_{base_dif_orthoe_ret_bm} | 0.01 | 0.06 |
| k_{k_dif_orthoe_ret_bm} | 0.01 | 0.06 |
| k_{k_migr_ery_pl} | 0.01 | 0.06 |
| k_{kbase_amo_bm} | 0.005 | 0.1 |
| k_{k_apo_cfu_e_bm} | 0.0001 | 0.001 |
| k_{k_apo_cfu_e_preproe_bm} | 0.0001 | 0.0001 |
| k_{IC50_epi_cfu_e} | 0.1 | 1    |
| k_{Imax_epi_cfu_e} | 0.0001 | 0.01 |
| k_{Id_epoR}    | 10   | 100  |
| k_{gamma_hb}  | 1    | 10   |
| k_{gamma_hif1a} | 1 | 5    |
| k_{coef_dif_epi} | 100 | 2000 |
| k_{kbase_apo_cfu_e_bm} | 0.01 | 0.2 |
| k_{kbase_dif_cfu_e_preproe_bm} | 0.008 | 0.03 |

*min* minimal value, *max* maximal value
Approach 1: Monte Carlo Markov chain

Let us consider in more detail the methodology of Approach 1. The main idea of this method is based on the Bayesian paradigm. First, we chose a set of \( M \) variable parameters \((p_1,\ldots,p_M)\) used in the model, for example, the ones to which the control function is the most sensitive. In the erythropoiesis model in the current study, \( M = 39 \). Subsequently, we assumed that the prior distributions of these parameters are described by the truncated normal distribution for the parameters \((\mu_i, \sigma_i), i = 1,\ldots,M\), where \( \mu_i \) is fixed on the level fitted in the model for the “reference patient”, \( \sigma_i = \frac{\mu_i}{2} \).

Let us denote the dependence between the vector of the control function and parameter values by \( y = y(p_1,\ldots,p_M) \). In our case, \( y \) relates to the RET_p plnorm. The task was to generate a series of vectors \((p^{(1)}_1,\ldots,p^{(1)}_M),\ldots,(p^{(N)}_1,\ldots,p^{(N)}_M)\), such that the sample mean and SD of the control function

\[
y^{(1)} = y(p^{(1)}_1,\ldots,p^{(1)}_M),\ldots,y^{(N)} = y(p^{(N)}_1,\ldots,p^{(N)}_M)
\]

are close to the observed \( m \) and \( sd \), where \( N \) is the number of patients in clinical trials. We used the MCMC approach to generate such a set of vectors. The stages of the algorithm are as follows:

1. First, generate the values of \( p_1,\ldots,p_M: (p^{(1)}_1,\ldots,p^{(1)}_M) \) using data from prior distributions.
2. Calculate the control function value \( y^{(1)} = y(p^{(1)}_1,\ldots,p^{(1)}_M) \).
3. Calculate the likelihood \( L^{(1)} \) for \( y^{(1)} \) with respect to truncated normal distributions with parameters \( m \) and \( sd \) using the following formula:

\[
L^{(1)} = \prod_{i=1}^{T} \frac{1}{z_i \cdot sd_i} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(y^{(1)} - m_i)^2}{2sd_i^2}}, \text{ where}
\]

\[
z_i = 1 - F(a_i), \quad F(a_i) = \int_{-\infty}^{a_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,
\]

\[
a_i = -\frac{m_i}{sd_i}
\]

4. Calculate the likelihood \( p^{(1)} \) for \((p^{(1)}_1,\ldots,p^{(1)}_M)\) with respect to the independent prior distributions of the parameters using the following expression:

\[
p^{(1)} = \prod_{i=1}^{M} \frac{1}{z_i \cdot \sigma_i} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x^{(1)} - m_i)^2}{2\sigma_i^2}}, \text{ where}
\]

\[
z_i = 1 - F(a_i), \quad F(a_i) = \int_{-\infty}^{a_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,
\]

\[
a_i = -\frac{m_i}{\sigma_i}
\]

5. The overall likelihood is the product \( L^{(1)} p^{(1)} \).
6. Generate proposal for \((p^{(2)}_1,\ldots,p^{(2)}_M) \) —the randomly chosen vector from some distribution with mean in \((p^{(1)}_1,\ldots,p^{(1)}_M)\).
7. Repeat steps 2–5 to obtain \( L^{(prop)} p^{(prop)} \).
8. If \( L^{(prop)} p^{(prop)} > L^{(1)} p^{(1)} \) proposal is accepted, then

\[
p^{(2)}_1 := p^{(prop)}_1\ldots p^{(2)}_M := p^{(prop)}_M
\]

and a new proposal is generated from the distribution with mean in the range of \((p^{(2)}_1,\ldots,p^{(2)}_M)\) (go to step 6).

9. Repeat steps 6–8 till \( N \) accepted vectors \((p^{(1)}_1,\ldots,p^{(1)}_M),\ldots,(p^{(N)}_1,\ldots,p^{(N)}_M)\) are obtained.

Thus, we obtained \( N \) vectors of variable parameters and then simulated our model with them and obtained \( N \) values of the control function for each time point. As \( n_{MCMC} \) and \( sd_{MCMC} \) are the approximations for \( m \) and \( sd \) for each time point, the set vectors \((p^{(1)}_1,\ldots,p^{(1)}_M),\ldots,(p^{(N)}_1,\ldots,p^{(N)}_M)\) represent the population of \( N \) VPs.
Approach 2: model fitting to Monte Carlo sample

In this section, we explain the details of Approach 2. The main idea of this method is to fit the variable parameters \( p_1, \ldots, p_M (M = 39) \) to pseudo-experimental data that were randomly sampled from distributions with the same parameters \((m, \ t, \ t = 1, \ldots, T, T = 16)\) as the experimental data at each time point.

Similar to the previous section, let us denote the dependence between the vector of the control function and parameter values by \( y = y(p_1, \ldots, p_M) \). The first stage of the methodology proposed here assumes the generation of a pseudo-experimental dataset of vectors \( \bar{y}^{(1)}, \ldots, \bar{y}^{(N)} \) characterized by a series of means \((m = (m_1, \ldots, m_T))\) and a series of SDs \((sd = (sd_1, \ldots, sd_T))\) equal to those in the experimental data. After the variable parameters are fitted to each of the pseudo-experimental vectors \( \bar{y}^{(1)}, \ldots, \bar{y}^{(N)} \), a series of vectors \((p_1^{(1)}, \ldots, p_M^{(1)}), \ldots, (p_1^{(N)}, \ldots, p_M^{(N)})\) are obtained, where \( N \) is the number of patients in a clinical trial. The stages of the algorithm are as follows:

(1) Generate sets of pseudo-experimental data from an appropriate log-normal distribution with parameters \( m = (m_1, \ldots, m_T) \) and \( sd = (sd_1, \ldots, sd_T) \) obtained from the experimental data for each time point.
\[
\bar{y} = \left( y_1^{(1)}, \ldots, y_1^{(N)}, \ldots, y_T^{(1)}, \ldots, y_T^{(N)} \right), T = 16.
\]

(2) Set physiological ranges for the variable parameters.

(3) Fit the variable parameters (to fit the parameters we used “L-BFGS-B” method in R) to minimize the following objective function:
\[
\sum_{i=1}^{T} \sum_{j=1}^{N} \left( y_j^{(i)} - \hat{y}_j^{(i)} \right)^2,
\]

where \( (y_1^{(1)}, \ldots, y_T^{(1)}) = y(p_1^{(1)}, \ldots, p_M^{(1)}) \)

The series of fitted parameter vectors \((p_1^{(1)}, \ldots, p_M^{(1)}), \ldots, (p_1^{(N)}, \ldots, p_M^{(N)})\) represent the required population of \( N \) VPs, as obtained by the averages for \( i \) th mean and sd values for each time point \( m^{\text{fitting}} = (m_1^{\text{fitting}}, \ldots, m_T^{\text{fitting}}) \) and \( sd^{\text{fitting}} = (sd_1^{\text{fitting}}, \ldots, sd_T^{\text{fitting}}) \) of \( (y_1^{(i)}, \ldots, y_T^{(i)}), i = 1, \ldots, N \) are the approximations for \( m \) and \( sd \).

Approach 3: Population of clones

The main idea of this method is to propose a series of patient groups. Each group is described by its own control function value, i.e., patients within groups are identical, whereas patients from different groups are described by different control function values and parameters vectors respectively. To obtain correspondence with the data we calculate the relative group sizes in terms of fractions of patients in the population relating to the particular group. By doing this we ensure that the overall control function distribution had the same values of \( m \) and \( sd \) as those in the experimental data.

In the first stage, we selected one time point for further analysis by calculus. We denoted this moment as \( t_0 \). In particular, in the case of our model \( t_0 = t_8 \). To provide a robustness check we analyzed the results using different time points and find out that the choice of time point does not significantly affect the quality of the results. However, as the results are slightly dependent on the time point chosen, we recommend to create VPs using different time points (the method is computationally very fast, so it is not a problem to create such VPs) and choose the one providing the best fit over the whole time range. After that, for simplicity, we chose the number of groups \( L = 5 \) (however, it can be of any size, not necessary \( 5 \)) control function values from the control function variation range for this time point \((y_0^{(1)}, \ldots, y_0^{(L)})\), as it was approximately uniformly located between the \( m \) and \( m \pm sd \) values, as shown in Fig. 3. The stages of this algorithm are as follows:

(1) Fit variable model parameters \( p_1, \ldots, p_M (M = 39) \) to these control function values to obtain \( L \) different sets of parameters describing different patient groups: \( y_0^{(1)}, \ldots, y_0^{(L)} \).

(2) Assume that at time \( t_0 \), the control function has a trapezoidal distribution (Fig. 3).

(3) Further, based on the definition of mathematical expectation, SD, and similarity of triangles, we composed the following system of equations:
\[
N_0^{(l)} = \frac{Nl}{N},\text{where}\ N\text{is number of patients in a clinical trial}, l = 1, \ldots, L, L = 5
\]

\[
n_0^{(1)} + \ldots + n_0^{(L)} = 1
\]

\[
sd_0 = \frac{N_0^{(1)}(y_0^{(1)} - m_0) + \ldots + N_0^{(L)}(y_0^{(L)} - m_0)}{N - 1}
\]

\[
m_0 = n_0^{(1)}y_0^{(1)} + \ldots + n_0^{(L)}y_0^{(L)}
\]

\[
n_0^{(1)} - n_0^{(1)} = y_0^{(1)} - m_0
\]

\[
n_0^{(2)} - n_0^{(2)} = y_0^{(2)} - y_0^{(1)}
\]

\[
n_0^{(3)} - n_0^{(3)} = y_0^{(3)} - y_0^{(2)}
\]

\[
n_0^{(4)} - n_0^{(4)} = y_0^{(4)} - y_0^{(3)}
\]

(4) Solving the system of linear equations, we obtained the necessary fractions of VPs: \( n_0^{(1)}, \ldots, n_0^{(L)} \).

Therefore, our VP is represented by the \( L = 5 \)
groups of patients and in the group \( l \) we have \( n_i^{(l)}N \) patients (\( l = 1, \ldots, L \)).

(5) After that, using corresponding \( L = 5 \) parameters vectors describing typical patient in each of the group, we calculate control function values for other time points for each of the groups.

(6) Using the obtained group sizes (fractions \( n_i^{(l)}, \ldots, n_i^{(L)} \)) we calculate the means and standard deviations of the control function over the whole population for each point in time:

(a) Mean values for each time point are calculated using the following formula:

\[
    m_t = \sum_{i=1}^{L} n_i^{(l)} \cdot y_t^{(l)}, \quad t = 1, \ldots, 16
\]

(b) Standard deviation values for each time point are calculated as follows:

\[
    sd_t = \sqrt{\frac{1}{\sum_{i=1}^{L} n_i^{(l)} (y_t^{(l)} - m_t)^2}}, \quad t = 1, \ldots, 16
\]

**Approach 4: Stochastically bounded selection**

The main idea of this method was to select a sample of \( N \) (where, \( N \) is the number of patients in a clinical trial) elements from a relatively large set of random vectors of variable parameters \( p_1, \ldots, p_M (M = 39) \) such that they constitute a subpopulation of the initial VP population describing the results of a clinical trial.

Similar to Approaches 1 and 2, let us denote the dependence between the vector of the control function and parameter values by \( y = y(p_1, \ldots, p_M) \). First, in case of the example considered below, we generated a large series of vectors of variable parameters: \( (p_1^{(1)}, \ldots, p_M^{(1)}), \ldots, (p_1^{(S)}, \ldots, p_M^{(S)}) (S = 4000) \). We used the continuous uniform distribution for the ranges indicated in Table 4. The value of \( S \) depends on the model complexity and the available computing power. The larger this value is, the more accurate the estimations obtained are; however, the time needed to obtain the initial random sample is larger too.

After the vectors for variable parameters were generated, we simulated model with this vectors and as the result of these simulations we obtained a large series of vectors of control functions: \( y : (y_1^{(1)}, \ldots, y_J^{(1)}), \ldots, (y_1^{(S)}, \ldots, y_J^{(S)}) \), \( T = 16 \).

The stages of the algorithm are as follows:

1. Find the minimal \( Y_{min}^t \) and maximal \( Y_{max}^t \) value of the generated control function sample at each time point \( t \in [1, \ldots, T] \).
2. Generate \( N \) random values for \( y \) from truncated lognormal distribution [13] with the following characteristics:

   (1) \( m = (m_1, \ldots, m_T), \quad sd = (sd_1, \ldots, sd_T) \), and \( T = 16 \) are the same as in experimental data

   (2) distribution is concentrated in the interval \([Y_{min}^t, Y_{max}^t]\)

   Therefore, we obtain the following sets of vectors of the control function:

\[
    \hat{y} : (\hat{y}_1^{(1)}, \ldots, \hat{y}_T^{(1)}), \ldots, (\hat{y}_1^{(S)}, \ldots, \hat{y}_T^{(S)})
\]

3. Choose one time point for further fitting. In the current study, we did not propose a formal criterion for the choosing the time point. To obtain the results, we tried to use different time points and then chose the one that provided the best fit in visual assessment.

4. For each of the \( N \) values of \( \hat{y}_0 : \hat{y}_0^{(1)}, \ldots, \hat{y}_0^{(S)} \) find the nearest value of the control function in the initial sample \( y_0 : y_0^{(1)}, \ldots, y_0^{(S)} \) and select the corresponding values for all time points in accordance with the vector.

Thus, the parameters corresponding to the control function values observed in step 4 describe the VP.

**Global sensitivity analysis**

We performed a global sensitivity analysis to assess the impact of each variable parameter on the dynamics of the model. We used the partial rank correlation coefficient

![Image](Image 184x626 to 544x732)
method with Pearson correlation coefficients. A thorough description of this method has been provided in a paper by [14]. The partial rank correlation coefficient values were in the range of $-1$ to $1$. Values close to zero indicate that the parameter was not sensitive. The results of the sensitivity analysis are provided in the Results section.

**Results**

In this section, we compare the aforementioned approaches for VP population generation. The comparison was made for a relatively small (5 patients) and large population size (207 patients). In the first case, the comparison is qualitative, whereas in the second case, additional rigorous criteria were also considered.

**Comparison of the approaches in the case of five patients**

As the target sample size was extremely small, the performance of the approaches could be compared by visual assessment only. We want to note, that Approach 3 was inapplicable in case of a small sample size. This follows from the fact that Approach 3 is based on the calculation of group sizes. This sizes are formulated in terms of fractions of patients in the population. When the population is small (i.e. in includes only a few patients) the description in terms of fractions is evidently inapplicable (imagine, what happens when we take 10% from five people). Figure 4 shows the results obtained using the 3 approaches in case of five patients.

In Fig. 4a, we present the mean and SD values of the control function $RET_{pl\_norm}$ distribution predicted in Approach 1 at every time point and those for the experimental values. Similarly, in Fig. 4b, c, we show the characteristics of the predicted control function distributions and the corresponding experimental values at every time point for Approaches 2 and 4, respectively.

Figure 4a shows the poor precision of the mean and SD estimates in Approach 1. This was not surprising, as a large sample is required to obtain reliable results in the MCMC approach. Better results were obtained in Approach 2 (Fig. 4b) as it predicted the mean values relatively accurately; however, the SD estimates were too small. This flaw was overcome in Approach 4 (Fig. 4c), in which relatively good estimates of both mean and SD values were obtained, despite the fact that the sample size was extremely small. Therefore, Approach 4 can be considered as the preferred approach.

**Comparison of the approaches in the case of 207 patients**

Besides visual assessment, in the case of a relatively large sample size, the approaches can be compared more rigorously. To do this, we used the one-sample $t$-test and QQ plot. In the case of Approach 3, it is not possible to apply comparisons in terms of QQ plot and one-sample $t$-test,
because the final sample consists of only five unique types of virtual patients.

**Visual assessment**

Figure 5 shows the results obtained by the four approaches for 207 patients. In Fig. 5a–d, we present the results of Approaches 1, 2, 3, and 4, respectively. Similar to Fig. 4a–c, Fig. 5a–d depict the mean and SD values of the predicted control function distributions and the corresponding experimental values at every time point. The figure clearly indicates that all the approaches accurately predict the mean values at every time point; however, relatively precise estimates of SD were only obtained in Approach 4.

**QQ plot**

Using a QQ plot, two distributions can be compared based on their respective quantile positions. Therefore, we compared the predicted distributions with the experimental distributions for each approach, except for Approach 3. Figures of the Online Resource 1 show the QQ plots for each approach at every time point. In the main text of the study we presented only the qualitatively good-fit examples (i.e. the ones with the closest position to the bisector).

Figure 6 shows QQ plots for the 3 time points at which the predicted and experimental distributions coincide in Approach 1. Figures 7 and 8 illustrate the best time points for Approaches 2 and 4, respectively. Approach 4 provided the largest number of relatively accurate time points.

**One sample t-test**

Since the experimental data are given in the form of a series of mean, \( m = (m_1, \ldots, m_T) \), and SD, \( sd = (sd_1, \ldots, sd_T) \), values, we used one-sample t-test to assess their statistical significance. The p-values generated for each approach, except Approach 3, are shown in Table 2. The data in the table shows that in case of Approach 1, 2, and 4, statistical significance is achieved at 3, 2, and 4 time points, respectively (highlighted by italics in Table 2). Therefore, with respect to this criterion, Approach 4 is the preferred approach.
This section presents the analysis of the distributions obtained for the variable parameters (final VP sample) in each approach, except for Approach 3, because the final sample, obtained in this approach, consists of only five unique types of virtual patients. In Fig. 9 the results of the global sensitivity analysis for all 39 model parameters are presented. This analysis was conducted using the partial rank correlation coefficient method [14]. From the list of parameters with the highest sensitivity we selected five most important in view of Erythropoiesis process. These parameters are: \( k_{\text{dif_ret_ery_pl}} \), \( k_{\text{apo_ery_pl}} \), \( \text{IC50_a-po_epo_cfu_e} \), \( \text{EC50_pro_epo_cfu_e} \), and \( \gamma_{\text{hif1a}} \).

Sensitivity analysis revealed that the response to EPO is sensitive to its effect on apoptosis, CFU-erythroid cell proliferation, and EPO synthesis-related processes. It has been noted that CFU-erythroid cells are absolutely dependent on EPO as they carry the highest EPO receptor density; therefore, they are the most sensitive to EPO among other erythroid progenitors. Parameters \( k_{\text{dif_ret_ery_pl}} \) and \( k_{\text{apo_ery_pl}} \) directly influence erythrocyte count, which contributes to hemoglobin much more than reticulocytes. Hemoglobin, being an oxygen carrier in turn affects HIF, due to which EPO production is regulated and EPO levels return to the steady-state. Parameter \( \gamma_{\text{hif1a}} \) as a power-law indicator of the rate of HIF-dependent synthesis of EPO mRNA strengthens the feedback loop between EPO level and erythrocyte count non-linearly.

The sample distributions for the top five sensitive parameters in the results of each approach are shown in Fig. 10. Figure 10a depicts the results of Approach 1. The distributions shown here have unimodal shapes, which is not surprising as prior distributions in the MCMC algorithm are truncated normal distributions. Figure 10b illustrates the parameter distributions in Approach 2. As we used a log-normal distribution for the generation of the pseudo experimental data, the shapes of the resulting parameter distributions inherit some specific features of this distribution type. However, the existence of the upper and lower bounds resulted in the appearance of the second mode. Figure 10c shows the resulting parameter distributions for Approach 4. In this case, the distribution shapes were more uniform, which is explained by the uniform initial parameter distributions in this approach.
In this section, we discuss the advantages and disadvantages of the approaches and describe probable directions of future research. This study aimed to compare different techniques to generate VP populations based on experimentally measured mean data and statistics. We proposed, implemented, and compared four different approaches to generate VP populations. These are (1) MCMC, (2) model fitting to Monte Carlo sample, (3) population of clones, and (4) stochastically bounded selection. For the purpose of the comparison of these approaches, we used the QSP model of erythropoiesis. To generate the VP populations, 39 parameters of the model were chosen. To ensure that the appropriate VP populations were selected, clinical data describing the changes following erythropoietin administration in the reticulocyte count in the blood of five patients were used. The main purpose of this study was to obtain the distribution of the control function described by the same number of virtual patients as those in the clinical trial. However, the sample of five patients was too small to represent a reliable quantitative comparison. Therefore, a larger VP population of 207 patients was generated. The comparison showed that all the approaches proposed are potentially applicable in different cases; therefore, we have discussed their advantages and disadvantages herein.

**Table 2** Comparison of the approaches in the case of 207 patients: one sample *t*-test

| Time, day | p-value      | Approach 1 | Approach 2 | Approach 4 |
|-----------|-------------|------------|------------|------------|
| 1         | 2.54E-06    | 8.83E-32   | 1.44E-05   |
| 2         | **9.14E-02**| 7.81E-26   | 6.39E-09   |
| 3         | 1.85E-10    | 1.82E-09   | 2.05E-37   |
| 4         | 3.95E-07    | 8.53E-09   | 2.29E-05   |
| 5         | 8.76E-08    | 0.0007822  | **0.2023** |
| 6         | 1.31E-08    | **0.5954** | **0.5695** |
| 7         | 6.08E-05    | 5.85E-14   | 1.73E-50   |
| 8         | 4.87E-09    | **0.1696** | 0.0005495  |
| 9         | **9.68E-01**| 2.71E-71   | 2.77E-115  |
| 10        | 4.54E-04    | 4.64E-24   | 2.77E-39   |
| 11        | 2.27E-04    | 3.47E-23   | 2.28E-18   |
| 12        | 1.22E-05    | 3.12E-07   | **5.68E-01** |
| 13        | 6.13E-05    | 1.94E-08   | **5.49E-02** |
| 14        | **9.39E-01**| 7.42E-105  | 2.10E-20   |
| 21        | 1.17E-02    | 4.69E-41   | 7.50E-34   |
| 28        | 2.79E-09    | 2.67E-12   | 1.29E-27   |

Bold values highlight the positions of the statistical distribution means match

**Fig. 8** Approach 4 QQ plots corresponding to the time points with good fit
The main disadvantage of Approach 1 is that it works for large samples only, i.e., if one needs to generate a small population of patients, the results of this technique will be unsatisfactory. Another disadvantage of this approach is that an appropriate SD estimate can be obtained only in cases where its value is relatively small. The algorithm of this approach also calls for the simulation of the model at each step. Therefore, if the model and number of patients are both large, this approach will be very time consuming.
The main advantage of Approach 2 is its potential accuracy as using this, the final population of patients is obtained by fitting the variable parameters to the generated pseudo experimental data having exactly the same mean and SD. However, this approach has a series of limitations, such as (1) only a limited number of parameters can be accurately fitted, (2) one can obtain precise fitting only in the case of small models, and (3) the physiological ranges of variable (fitted) parameters should be known.

The main advantage of Approach 3 is its computational simplicity. Moreover, regardless of the model complexity, a population having exactly the same mean and SD as in the data can be obtained. However, in the VP populations obtained using this approach, there is no intergroup variability, and the VP population size should be rather large.

Recall the three comparison criteria formulated in the introduction. To make the comparison explicit we condensed the described above approaches pros, cons and limitations into these three “dimensions”. These results are presented in Table 3.

From Table 3 it is evident that Approach 4 appears to be the most preferred with regard to the initially formulated requirements. As can be seen from the results presented in this study, Approach 4 appears to be the most universal as it allows the description of any number of patients from clinical trials, and it can be applied in the case of complex models with a large number of variable parameters.

However, one of the main disadvantages of Approaches 3 and 4 is that both the methods are based on the selection of virtual patients at a single time point only. On the one hand, this fact simplifies the methodology. However, most likely for this reason we obtain observable poor match for series of the time points. This limitation can be overcome through the automatic adjustment of other time points and by not using their experimental data (mean and SD for each time point). For the model used in this study, we consider such simplification to be acceptable, since (1) a comparison of the results of Approaches 3 and 4 with those of multi-point Approaches 1 and 2 shows that the results were even better; (2) when we work with small population sizes we are mostly interested in qualitative trends and not exact matchings.

Thereby, there are a series of natural improvements for Approach 4. Their introduction into the methodology will make Approach 4 even more universal and accurate. In the near future, we plan to add the following updates: (1) generation of VP populations describing the distributions of several control functions simultaneously, (2) introduction of the restrictions on biomarkers, and (3) selection of VP populations using all time points.

Based on the results of the global sensitivity analysis, we selected 5 out of 39 parameters, which are (1) described by one of the highest sensitivity and (2) are of interest from the Erythropoiesis process point of view. We use these parameters to analyze the resulting distributions of variable parameters in details. The histograms of these distributions show that the shapes of the parameter distributions significantly differed amongst the approaches. The boundaries of the physiological ranges of the variable parameters are listed in Table 2. A comparison of the final distributions of the variable parameters shown in Fig. 9 with the physiological ranges of the variable parameters (as shown in Table 2) revealed that the final distributions of the parameters mostly belong to the corresponding physiological boundaries.

### Conclusions

The paper presents four different approaches to generate virtual patient populations based on experimentally measured statistics. Comparison of the results of application of the four approaches to populations of two different sizes shows that the stochastically bounded selection is the preferable one in both cases. Moreover, this approach appears to have the smallest computational complexity and, therefore, it can be applied even to the high dimensional models with large number of variable parameters.

### Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s10928-022-09814-y.

### Acknowledgements

We are grateful to the two anonymous reviewers for their insightful comments and suggestions which allowed us to significantly improve the results presentation.

### Author contributions

GK: created methodology and performed all data analysis; OD and GK: discussed and analyzed the results; AS and GL: created model of erythropoiesis. All authors drew the main conclusions and wrote the manuscript.
Declarations

Conflict of interest Authors declare that there are no relevant financial or non-financial competing interests to report.

References

1. Chien JY, Friedrich S, Heathman MA, de Alwis DP, Sinha V (2005) Pharmacokinetics/Pharmacodynamics and the stages of drug development: role of modeling and simulation. The AAPS J. https://doi.org/10.1208/aapsj070355
2. Knight-Schrijver VR, Chelliah V, Cucurull-Sanchez L, Le Novère N (2016) The promises of quantitative systems pharmacology modelling for drug development. Comput Str Biotechnol J. https://doi.org/10.1016/j.csbj.2016.09.002
3. Bloomingdale P, Karelina T, Cirit M, Muldoon SF, Baker J, McCarty WJ, Geerts H, Macha S (2021) Quantitative systems pharmacology in neuroscience: novel methodologies and technologies. CPT. https://doi.org/10.1002/psp4.12607
4. Chelliah V, Lazarou G, Bhatnagar S et al (2020) Quantitative systems pharmacology approaches for immuno-oncology: adding virtual patients to the development paradigm. Clinic Pharmacol Ther. https://doi.org/10.1002/cpt.1987
5. Dai W, Rao R, Sher A, Tania N, Musante CJ, Allen R (2020) A prototype qsp model of the immune response to SARS-CoV-2 for community development. CPT. https://doi.org/10.1002/psp4.12574
6. Balbas-Martinez V, Asin-Prieto E, Parra-Guillen ZP, Troconiz IF (2020) A quantitative systems pharmacology model for the key interleukins involved in Crohn’s disease. J Pharmacol Exp Ther. https://doi.org/10.1124/jpet.119.260539
7. Cheng Y, Thalhauser CJ, Smithline S, Pagidala J, Miladinov M, Vezina HE, Gupta M, Leil TA, Schmidt BJ (2017) QSP toolbox: computational implementation of integrated workflow components for deploying multi-scale mechanistic models. AAPS J. https://doi.org/10.1208/s12248-017-0100-x
8. Duffull S, Gulati A (2020) Potential issues with virtual populations when applied to nonlinear quantitative systems pharmacology models. CPT. https://doi.org/10.1002/psp4.12559
9. Schmidt BJ, Casey FP, Paterson T, Chan JR (2013) Alternate virtual populations elucidate the type I interferon signature predictive of the response to rituximab in rheumatoid arthritis. BMC Bioinform. https://doi.org/10.1186/1471-2105-14-221
10. Rieger TR, Allen RJ, Bystricky L, Chen Y, Colopy GW, Cui Y, Gonzalez A, Liu Y, White RD, Everett RA, Banks HT, Musante CJ (2018) Improving the generation and selection of virtual populations in quantitative systems pharmacology models. Prog Biophys Mol Biol. https://doi.org/10.1016/j.pbiomolbio.2018.06.002
11. Alexander Stepamov, Galina Lebedeva (2020) Stimulation of erythropoiesis with ESA or blood donation: QSP model, ACoP11.;2688–3953, Vol 2
12. Ramakrishnan R, Cheung WK, Wacholtz MC, Minton N, Jusko WJ (2004) Pharmacokinetic and pharmacodynamic modeling of recombinant human erythropoietin after single and multiple doses in healthy volunteers. J Clinic Pharmacology. https://doi.org/10.1177/009127004268411
13. Zaninetti L (2017) A left and right truncated lognormal distribution for the stars. arXiv. https://doi.org/10.22606/adap.2017.23005
14. Zhang X-Y, Trame MN, Lesko LJ, Schmidt S (2015) Sobol sensitivity analysis: a tool to guide the development and evaluation of systems pharmacology models. CPT. https://doi.org/10.1002/psp4.6

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.