

ORIGINAL ARTICLE

Efficient Generation and Selection of Virtual Populations in Quantitative Systems Pharmacology Models

RJ Allen*, TR Rieger and CJ Musante

Quantitative systems pharmacology models mechanistically describe a biological system and the effect of drug treatment on system behavior. Because these models rarely are identifiable from the available data, the uncertainty in physiological parameters may be sampled to create alternative parameterizations of the model, sometimes termed “virtual patients.” In order to reproduce the statistics of a clinical population, virtual patients are often weighted to form a virtual population that reflects the baseline characteristics of the clinical cohort. Here we introduce a novel technique to efficiently generate virtual patients and, from this ensemble, demonstrate how to select a virtual population that matches the observed data without the need for weighting. This approach improves confidence in model predictions by mitigating the risk that spurious virtual patients become overrepresented in virtual populations.

CPT Pharmacometrics Syst. Pharmacol. (2016) 5, 140–146; doi:10.1002/psp4.12063; published online 17 March 2016.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC? Parameter uncertainty in quantitative systems pharmacology models may be explored by the creation of virtual patients, which are typically weighted to form virtual populations to match clinical populations. Several algorithms designed to weight virtual patients have previously been published. • WHAT QUESTION DID THIS STUDY ADDRESS? Given that the parameters of a systems model are underconstrained, can we explore this uncertainty to efficiently generate physiologically reasonable patients and construct virtual populations where weighting is not necessary? • WHAT THIS STUDY ADDS TO OUR KNOWLEDGE This study outlines a methodology that improves on previous methods for efficiently generating and selecting virtual patients to match clinical population-level statistics. The final fitted populations will closely match empirical data, with all virtual patients weighted equally, which avoids the potential for overweighting certain solutions and skewing simulation results found in some previous algorithms. • HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS Generation of realistic virtual populations, and a deeper exploration of parameter uncertainty, should lead to better confidence in the predictions and better quantification of uncertainty of systems pharmacology models, particularly in the context of clinical trial simulations and analysis.

Quantitative Systems Pharmacology (QSP) models are an effective approach for gaining mechanistic insight into the complex dynamics of biological systems in response to drug treatment.^{1–3} QSP models in the drug discovery and development process have been utilized for increased confidence in rationale for early development targets, preclinical to clinical translation, and predictions of clinical response to novel therapeutics. To be fit for this purpose, these models must include sufficient biological scope and mechanistic detail to link pathway modulation to overall system response.^{4–8} Due to the complexity of the biology, the iterative model-building process frequently results in a model that is a large, nonlinear, multiscale system of equations. Many different data sources are required to quantify QSP models, including *in vitro* and *in vivo* preclinical and clinical data; moreover, the resulting models are frequently underconstrained by any one dataset.⁹ Therefore, to explore the impact of known variability and uncertainty,^{10–14} QSP models are simulated using ensembles of parameterizations often termed “virtual patients” or “VPs.” A virtual population (VPop) that reflects individual subject and population-level

characteristics of a typical clinical cohort provides increased confidence that prospective simulations of response to novel therapeutics will reflect the intersubject variability seen in the clinic, and may help to identify responders and nonresponders to treatment.

Ensembles of VPs are often sufficient for exploring the broad range of responses that are possible from perturbing a model (pharmacologically or otherwise), but the outcome will not necessarily reflect the distribution (e.g., log-normal) of population-level data.¹⁵ The result is a range of predictions from the model, which are all possible outcomes but fail to provide insight into the probability of observing that outcome in a clinical trial. Previous authors have overcome this critique by weighting model outputs,¹³ or model components,¹⁴ to create VPops.

Klinke proposed linearly weighting each VP, with some receiving a weight greater than 1/N (where N is the number of VPs in the ensemble) so that the mean and standard deviation of the VPop match the desired population characteristics.¹³ This approach is intuitive and easily implemented, but this also can be computationally expensive,

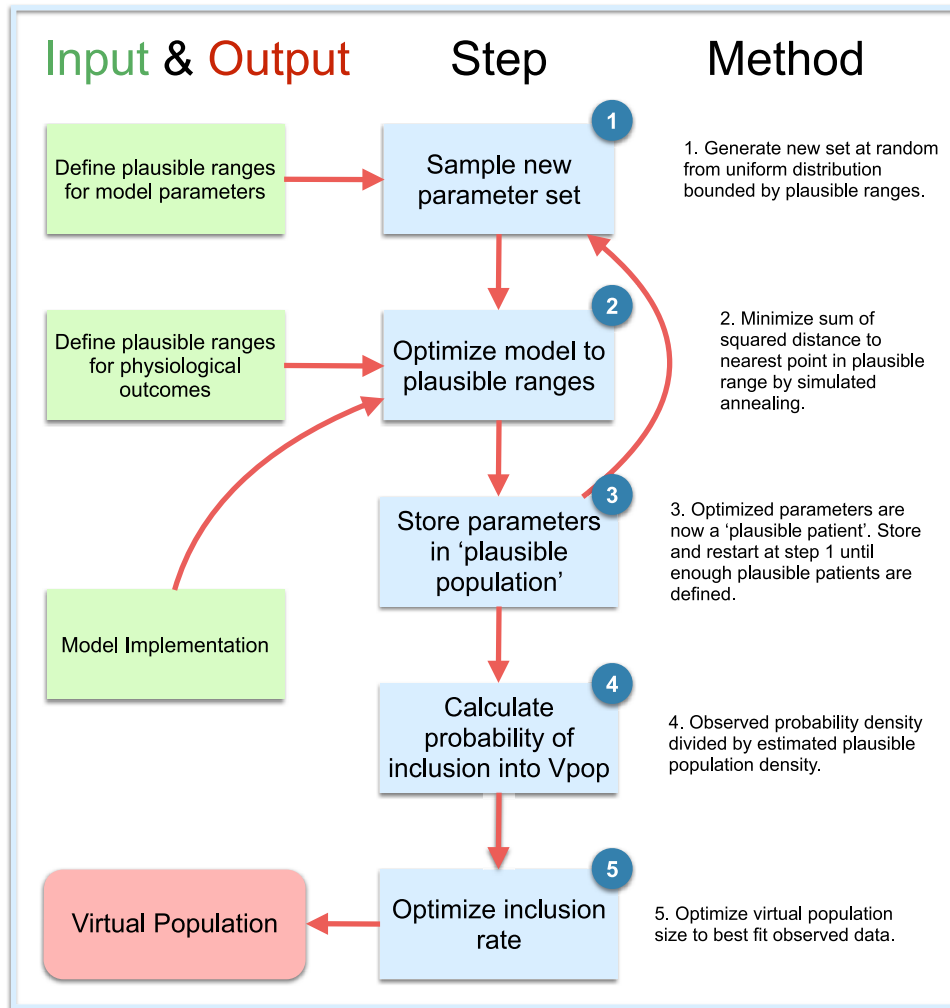


Figure 1 Overview of algorithm for efficient generation and prevalence-based selection of virtual patients. To generate virtual patients from a model, the prior information (green boxes) is used to define physiologically reasonable ranges for model outputs and parameter values. An initial parameter guess is optimized until model outputs are physiologically plausible. This is repeated multiple times to form a plausible population. A virtual population is constructed by selecting from this population with probability proportional to the prevalence in the real population relative to the prevalence in the plausible population. This selection is optimized to produce the best virtual population given the patients in the plausible population.

requires refitting the VPop each time VPs are added or removed from the analysis, and can result in a dramatic overweighting of a few select VPs, which may skew the final simulation results. Schmidt *et al.* refined this approach by taking the weights off of the individual VPs and placing them on “mechanistic axes.”¹⁴ Their approach is computationally faster, allows new VPs to be incorporated into the VPop without refitting, and should avoid the problem of overweighting small numbers of VPs. However, this approach requires collecting parameters into mechanistic axes, possibly without *a priori* rationale, and in our experience is more challenging to communicate to a non-technical audience.

Here we propose a new algorithm for generating biologically reasonable VPop. We will show how this algorithm complements previous approaches by being intuitive, computationally efficient, and avoiding the problem of over-

weighting VPs. We demonstrate the utility of this new algorithm by fitting the joint distribution of low-density and high-density lipoproteins (LDL and HDL, respectively) from the National Health and Nutrition Examination Survey (NHANES)¹⁶ to a previously published model of lipoprotein metabolism by van de Pas *et al.*¹⁷

METHODS

A flow diagram of the algorithm is shown in **Figure 1**. To implement this procedure for a given model it is necessary to define bounds for input parameters and model outputs (e.g., steady states or dynamic behavior). If bounds cannot be defined empirically, feasible ranges of parameter values can be asserted from physiological knowledge or theoretical considerations. For example, the tissue concentration of

Table 1 An overview of the terminology used in this article

Term	Definition	Attributes
Systems Pharmacology Model	$\dot{X} = f(X, t; p)$	X (species), and p(parameters) both vectors
Physiological Outcome	Any quantity that is calculated from the model that can be experimentally measured	i.e., $X_i(t)$, $dX_i(t)/dt$, $g(X)$...
Plausible Patient	A parameter set defining the model	Each parameter and physiological outcome is within biologically plausible ranges.
Plausible Population	A collection of Plausible Patients	None-specifically (all inherited from plausible patients)
Virtual Population	A subset of the Plausible Population	Distribution matches the physiological outcomes for which we have such information.
Virtual Patient (VP)	A Plausible Patient that is also in the virtual population	Parameters and physiological outcomes in plausible ranges. Probability of observing set of observations in VP approximates probability of observations in real patient.

a species may not be known, but typical weight and water content of that tissue may be known, which allows us to put an upper limit on the species concentration.

We have provided a detailed description of terminology, definitions, and the derivation of this algorithm in **Table 1** and the **Supplementary Material**. Briefly, our approach is to generate a large number of “plausible patients.” We define these patients as a parameter set for which every component of the model (whether it be the parameter values themselves, computed species concentrations, or combinations of these that are experimentally measurable) falls into a biologically plausible range. From this “plausible population” we can then select the virtual population such that it matches the empirical distribution of interest. This is achieved by calculating a probability of inclusion of a plausible patient into the virtual population. This probability is computed from both the empirical distribution and the density of plausible patients (see **Supplementary Materials** for more details).

An important prerequisite to this approach is the ability to generate a large number of plausible patients within the region of the empirical data. To accelerate this process we take an initial parameter guess (within the predefined bounds) and optimize this choice until the required outputs are within physiologically plausible ranges. Rather than

optimize to specific points, it is more efficient to be agnostic as to where in the plausible ranges the optimization routine ends. To implement this we shift the typical cost function $f(p)$ we would use optimizing a model to a new function, $g(p)$, where we consider both as purely dependent on the parameter set p . If we constrain parameters using a number of model outputs $M_i(p)$, with data d_i , then f (in the simplest, unweighted case) would be:

$$f(p) = \sum (M_i(p) - d_i)^2.$$

To generate plausible patients, we modify this sum-of-squared errors expression to:

$$g(p) = \sum_i \max \left[\left(M_i(p) - \frac{l_i + u_i}{2} \right)^2 - \left(\frac{u_i - l_i}{2} \right)^2, 0 \right]$$

where u_i and l_i are the predefined plausible upper and lower bounds, respectively, for $M_i(p)$. This expression ensures that if $M_i(p)$ is in the plausible range then the contribution of the corresponding term in the expression is zero. The effect of replacing $f(p)$ with $g(p)$ is visualized in 2D in **Figure 2**.

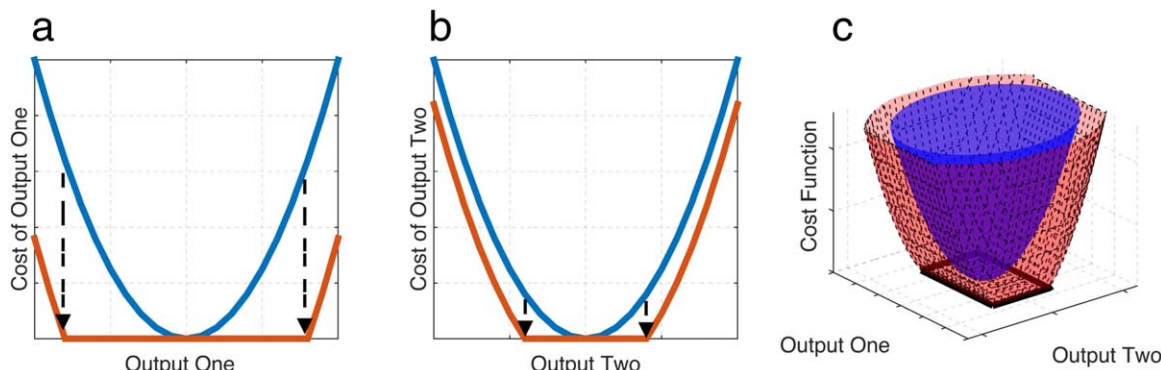


Figure 2 Cost function transformation for convergence to plausible virtual patients. Outputs of the model contribute to the cost function to be minimized by considering the sum of squared errors (SSE) from an associated experimental observation. For each observation we define a physiologically plausible range (arrows in **a,b**) and shift the SSE associated with that observation so that it is zero if the model output is in this range (**a,b**). Combining these transformations in each dimension leads to a broader cost function that is minimized by many points, rather than one (black rectangle in **c**).

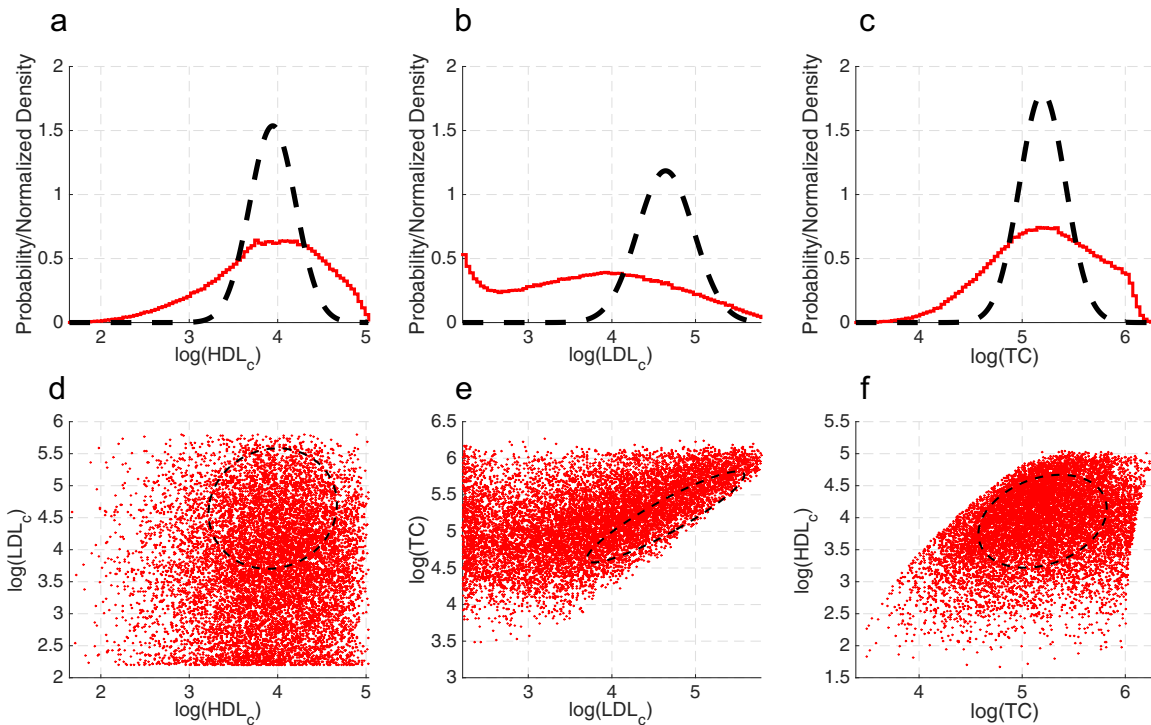


Figure 3 Comparison of the initial plausible population ($N=300,000$) with NHANES multivariate distribution ((a-c) black dotted lines estimated PDF, **Supplementary Figure 1a-c**). (d-f) 2D projection of the 95% confidence surface of the estimated probability density function).

To test this approach we used a previously published model of cholesterol metabolism.¹⁷ We chose this model because we could use publicly available data from the NHANES database¹⁶ to establish the empirical multivariate distribution for LDL cholesterol, HDL cholesterol, and total cholesterol (LDL_c, HDL_c, and TC, respectively). Note that the distribution of these variables is well approximated by a multivariate log-normal distribution (**Supplementary Figure 1**). For the remainder of the article we will describe these variables, either as model outputs or from NHANES, in log units (prior to taking the logarithm, units are mg/dL for cholesterol measures). The published version of this model does not explicitly calculate LDL_c or TC; instead, the outputs are HDL_c and non-HDL_c. From these two quantities TC is easily calculated. For full comparison with the NHANES data we introduced a new parameter to the model, k_{22} , which is simply the ratio between LDL_c and non-HDL_c. The supplied MATLAB (MathWorks, Natick, MA) file “input_ranges.m” gives details on parameter and output ranges for the van de Pas model. Also in the **Supplementary Material** is the code used in this case, which is easily modifiable for application to other models.

RESULTS

We generated $\sim 300,000$ plausible patients using the algorithm. As expected, the initial plausible population does not match the population-level statistics of the NHANES data

(**Figure 3**) but covers the empirical distribution (i.e., where there are likely to be empirical observations there are plausible patients).

We proceeded by calculating the probability of inclusion for each plausible patient. Once calculated, we established that most of the plausible patients are highly unlikely to be in the final distribution (**Figure 4**). This is due to the relative density of the plausible population to the empirical

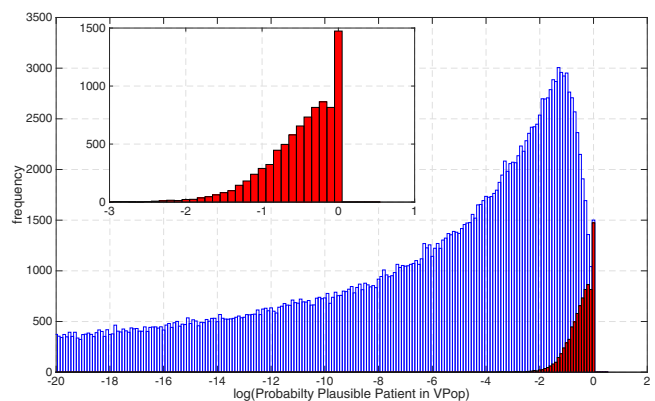


Figure 4 Histogram of plausible population selection probability. The probability of inclusion into the virtual population is calculated by optimized relative prevalence. The red histogram (main figure, and figure inset) is a virtual population that matches NHANES data, and is selected from the plausible population (blue histogram) based on displayed probability.

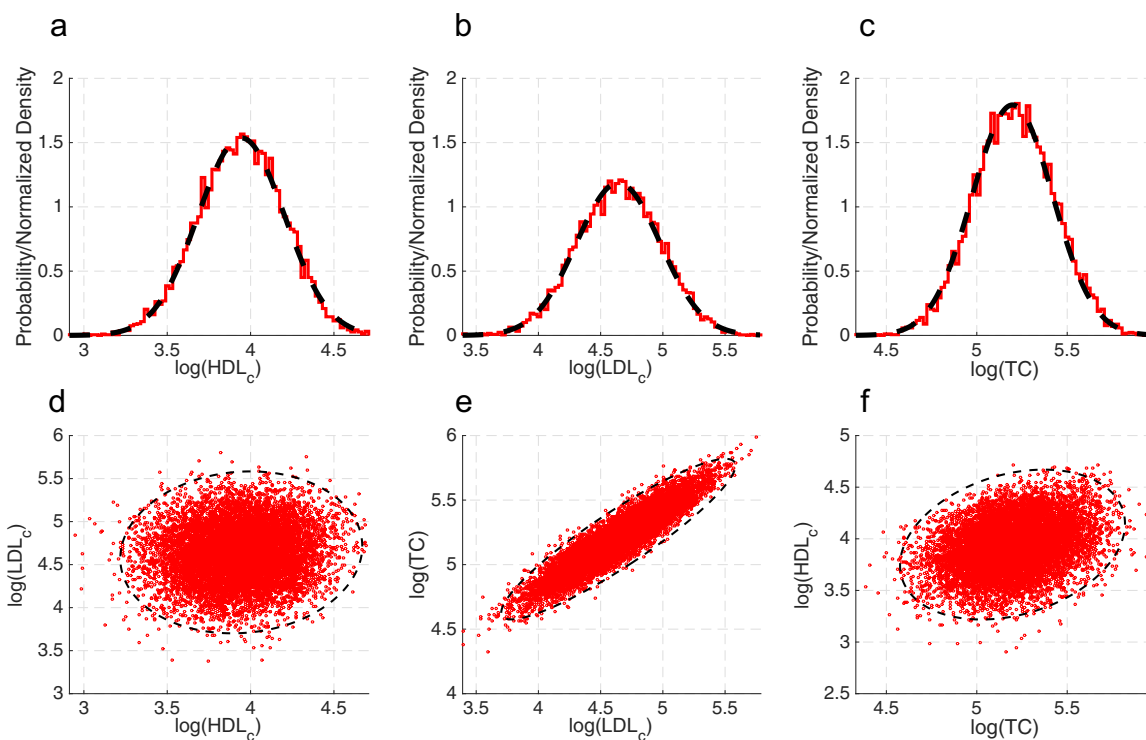


Figure 5 Comparison of a virtual population with NHANES multivariate distribution (dotted black lines). The virtual population (red dots and red histogram) matches the mean, variance, and covariance of the multivariate experimental distribution ((a-c) black dotted lines estimated probability density function, **Supplementary Figure 1a-c**. (d-f) 2D projection of the 95% confidence surface of the estimated PDF).

distribution. With these probabilities, only $\sim 2\%$ of the plausible population was selected to be in the virtual population (inset, **Figure 4**). Based on examining goodness-of-fit of the distributions it appears, in this case, there is no further value in increasing the size of the plausible population (**Supplementary Figure 2**).

The distribution of an example selection fits the NHANES data well (**Figure 5**). The 1D histograms (when normalized for comparison with the NHANES probability density function) are indistinguishable from the data (**Figure 5a-c**) and the correlations between variables also match the data based on visual predictive check.

When selecting a subset of VPs from a larger population, one concern is that the selected subset of VPs does not reflect the variability of the original ensemble, which was generated from the biologically plausible range of the parameters. Analyzing the final fitted population, we found little change in either the distribution of parameters or the correlation structure between the parameters (**Figure 6** and **Supplementary Figure 3**). This also shows that despite the virtual population being constrained against the NHANES data the parameter values of the virtual population (**Figure 6b**) are only slightly better constrained than those of the plausible population. Furthermore, correlations between parameters are only slightly increased in the virtual population (**Figure 6d**) vs. the plausible population (**Figure 6c**). At least in this case, constraining all outputs

into realistic ranges is a more stringent constraint than selection of a virtual population.

DISCUSSION

One of the primary uses for QSP models is to prospectively simulate the effects of a dynamic perturbation (pharmacological or otherwise) in populations of interest. Due to the underconstrained nature of these models it would be difficult to have confidence in the simulation results if we simulated a single parameterization of that model, even if that set of parameters is an excellent fit to the available data. For example, imagine creating a single hypercholesterolemia VP to simulate the effect of various anticholesterol therapies. For the baseline characteristics of the VPop, we have good data for the expected mean LDL and HDL (e.g., a prior clinical cohort), but we could still choose to mechanistically represent hypercholesterolemia several ways using the same model. We could increase cholesterol production, decrease clearance, or apply some combination of both. Our choice of how to parameterize that VP could have significant consequences for the sensitivity of the follow-on therapy simulations. Having impaired production vs. clearance of LDL could lead to differential responses to statins (production) vs. antiproprotein convertase subtilisin/kexin type 9 (clearance). A better approach is to explore the underconstrained nature of these models and sample

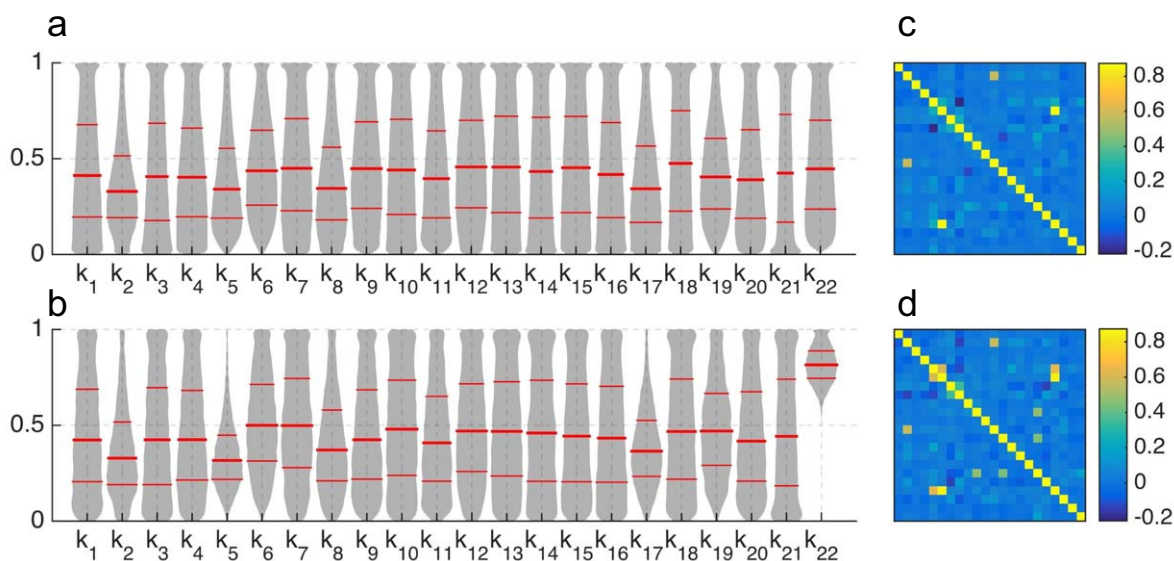


Figure 6 Degeneracy of the virtual population. Violin plots of the plausible and virtual populations (a,b, respectively) parameter values (normalized to each parameter's upper and lower bounds) and correlation matrix of the plausible and virtual population (c,d, respectively).

the biological uncertainty in the creation of plausible patients by varying mechanistic parameters, such as production and clearance rates, within biologically reasonable ranges. Simultaneously, we need to constrain the higher-level observables of the model based on known population distributions (e.g., the baseline characteristics of a clinical trial cohort).

An appealing aspect of the approach we outline is that, since the algorithm is probabilistic, once the plausible patients are generated any number of subpopulations can be selected as long as the generated patients reasonably cover the full range of the (sub)population. Additionally, for any particular population, any number of VPopS can be reselected to bootstrap the sensitivity of the model predictions to the choice of VPop.

Achieving an acceptable fit to the data distribution is only possible if the plausible patients densely cover the range of the observables. This method should be computationally tractable for many models. However, if we have higher-order density functions, we will likely require additional gains in efficiency, above and beyond what is presented here, in methods for generating sufficient plausible patients. One potential avenue, for a future iteration of this algorithm, may be to use methods that follow a directed search through the parameter space, such as using Markov Chain Monte Carlo (MCMC) algorithms.^{18–20} However, it should be noted that the method presented here is in fact a hybrid approach because the simulated annealing step, used to generate a plausible patient is essentially an MCMC method. The advantage of this approach is it generates plausible patients (and hence virtual patients) independently—which is critical for the purpose of making a virtual population. Also, once the plausible population is established new virtual populations, suitable for new applications, can be selected.

We believe that this method has broad utility in many cases. However, a limitation of our approach is the computational cost of generating large plausible populations, which may make this method unfeasible in models that are slow to simulate. It should be noted that prior methods^{13,14} generated smaller virtual populations in larger models, partly due to the computational cost of running the models. Nevertheless, the computational efficiency of a model does not alter the necessity for exploring the parameter space (which in this context equates to larger virtual populations). We therefore advocate for the optimization of large models for speed, such that a fuller exploration of parameter uncertainty is possible (either via the method presented here or alternatives). It is important to remark that the larger the model (specifically, in terms of the number of poorly constrained parameters) the greater the necessity for larger virtual populations to attempt to account for the uncertainty inherent in such a model.

An advantage of our technique over prior approaches is that it is relatively unbiased, while still leveraging all available information on possible parameter values. The approach by Klinke may overweight spurious model solutions, whereas the approach by Schmidt *et al.* requires binning of parameter values into mechanistic axis which, for axes containing more than one parameter, requires an assumption about the correlation between parameters in the population of interest that may not be supported by available data.

As an introduction to this algorithm, we demonstrated how to generate a VPop that matches the baseline characteristics of a population or clinical cohort; however, in practice, a dynamic model should be constrained additionally against as many in-scope perturbation experiments as possible (determined by available data). For this example, simulating changes in LDL_c and TC to standard-of-care lipid

therapies, such as statins and ezetimibe, would likely be an important step before using the model to predict the response to a novel mechanism. Ideally, information would be available detailing the distribution of the data before and after the application of therapy (i.e., not just summary statistics). The therapeutic response can be treated as a baseline constraint for VP selection, just as we used HDL_c and TC at baseline.

Future developments of this method will be driven by application to new models. We have applied this approach, without major adaptation, to an unpublished model of chronic kidney disease (42 Ordinary Differential Equations (ODEs), ~200 parameters) and to a model of body weight change⁶ (8 ODEs, ~50 parameters); we have made a brief summary of the results in these cases available.²¹ One challenge that we foresee, for virtual populations in general (i.e., not specific to this approach), is development of efficient ways to combine populations when merging distinct models. We believe that for two models that have large validated virtual populations, naïvely combining every possibility will be computationally daunting. Hence, we are interested in exploring more sophisticated approaches to this question.

Quantitative systems pharmacology models are becoming established as a valuable component of the drug discovery and development process. Communicating their complexity and uncertainty to an interdisciplinary project team is a critical but challenging component of their utility. Virtual populations are one tool that we have found to be successful in exploring mechanistic and parametric uncertainty in an intuitive framework that is easily understandable by most audiences. However, despite their widespread use, there are very few published methods for generating virtual patients and forming virtual populations. Here we have contributed an approach to efficiently generate virtual patients and construct a virtual population for which each patient is weighted equally. This method relies on the ability to generate large “plausible populations,” made up of plausible patients each of which is a candidate to become a virtual patient. Because a large plausible population is necessary, models that are slow to integrate (for example, with dynamics across multiple time-scales) may not be good candidates for this approach. However, in cases where quantitative predictions are required and the model is amenable to thorough examination of parameter space, we have found this method to be an improvement over previous approaches.

Conflict of Interest. The authors declare no conflicts of interest.

Author Contributions. R.A., T.R.R., and C.M. wrote the article; R.A. designed the research; R.A. and T.R.R. performed the research; R.A., and T.R.R. analyzed the data.

1. van der graaf, P.H. & Benson, N. Systems pharmacology: bridging systems biology and pharmacokinetics-pharmacodynamics (PKPD) in drug discovery and development. *Pharm. Res.* **28**, 1460–1464 (2011).
2. Allerheiligen, S.R.B. Next-generation model-based drug discovery and development: quantitative and systems pharmacology. *Clin. Pharmacol. Ther.* **88**, 135–137 (2010).
3. Sorger, P.K. *et al.* Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. A NIH White Paper by the QSP Workshop Group-October 2011 (2011).
4. Peterson, M.C. & Riggs, M.M. A physiologically based mathematical model of integrated calcium homeostasis and bone remodeling. *Bone* **46**, 49–63 (2010).
5. Hallow, K.M. *et al.* A model-based approach to investigating the pathophysiological mechanisms of hypertension and response to antihypertensive therapies: extending the Guyton model. *AJP Regul. Integr. Comp. Physiol.* **306**, R647–R662 (2014).
6. Hall, K.D. Predicting metabolic adaptation, body weight change, and energy intake in humans. *AJP Endocrinol. Metab.* **298**, E449–466 (2010).
7. Dalla Man, C., Rizza, R.A. & Cobelli, C. Meal simulation model of the glucose-insulin system. *IEEE Trans. Biomed. Eng.* **54**, 1740–1749 (2007).
8. Lu, J., Hübner, K., Nanjee, M.N., Brinton, E.A. & Mazer, N.A. An in-silico model of lipoprotein metabolism and kinetics for the evaluation of targets and biomarkers in the reverse cholesterol transport pathway. *PLoS Comput. Biol.* **10**, e1003509 (2014).
9. Gutenkunst, R.N. *et al.* Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**, e1891878 (2007).
10. Jamei, M., Dickinson, G.L. & Rostami-Hodjegan, A. A framework for assessing inter-individual variability in pharmacokinetics using virtual human populations and integrating general knowledge of physical chemistry, biology, anatomy, physiology and genetics: a tale of “bottom-up” vs “top-down” recognition of covariates. *Drug Metab. Pharmacokinet.* **24**, 53–75 (2009).
11. Valitalo, P.A.J. *et al.* Novel model-based dosing guidelines for gentamicin and tobramycin in preterm and term neonates. *J. Antimicrob. Chemother.* **70**, 2074–2077 (2015).
12. Brown, D. *et al.* Trauma in silico: Individual-specific mathematical models and virtual clinical populations. *Sci. Transl. Med.* **7**, 285ra61 (2015).
13. Klinke, D.J., II. Integrating epidemiological data into a mechanistic model of type 2 diabetes: validating the prevalence of virtual patients. *Ann. Biomed. Eng.* **36**, 321–334 (2008).
14. Schmidt, B.J., Casey, F.P., Paterson, T. & Chan, J.R. Alternate virtual populations elucidate the type I interferon signature predictive of the response to rituximab in rheumatoid arthritis. *BMC Bioinform.* **14**, 221 (2013).
15. van de Pas, N.C.A. *et al.* Predicting individual responses to pravastatin using a physiologically based kinetic model for plasma cholesterol concentrations. *J. Pharmacokin. Pharmacodyn.* **41**, 351–362 (2014).
16. NHANES Data. Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey Data. <<http://www.cdc.gov/nchs/nhanes.htm>> (2011–2012).
17. van de Pas, N.C.A., Woutersen, R.A., van Ommen, B., Rietjens, I.M.C.M. & de Graaf, A.A. A physiologically based in silico kinetic model predicting plasma cholesterol concentrations in humans. *J. Lipid Res.* **53**, 2734–2746 (2012).
18. Andrieu, C. & Thoms, J. A tutorial on adaptive MCMC. *Stat. Comput.* **18**, 343–373 (2011).
19. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
20. Peskun, P.H. Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–612 (1973).
21. Allen, R., Rieger, T. & Musante, C. Efficient generation of virtual populations of CKD patients and applications in quantitative systems pharmacology (presented at ACOP6). *F1000Research* **5**, (2016).

© 2016 The Authors CPT: Pharmacometrics & Systems Pharmacology published by Wiley Periodicals, Inc. on behalf of American Society for Clinical Pharmacology and Therapeutics. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Supplementary information accompanies this paper on the CPT: Pharmacometrics & Systems Pharmacology website (<http://www.wileyonlinelibrary.com/psp4>)