Received: 30 April 2024 | Revised: 23 August 2024 | Accepted: 10 October 2024

DOI: 10.1002/psp4.13270

ARTICLE

An integrated quantitative systems pharmacology virtual population approach for calibration with oncology efficacy endpoints

Nathan Braniff[1] | **Tanvi Joshi**[2] | **Tyler Cassidy**[3] | **Michael Trogdon**[1] | **Rukmini Kumar**[2] | **Kamrine Poels**[1] | **Richard Allen**[3] | **Cynthia J. Musante**[3] | **Blerta Shtylla**[1]

[1]Pharmacometrics & Systems Pharmacology, Pfizer Inc., La Jolla, California, USA

[2]Vantage Research Inc., Lewes, Delaware, USA

[3]Pharmacometrics & Systems Pharmacology, Pfizer Inc., Cambridge, Massachusetts, USA

Correspondence
Blerta Shtylla, 10777 Science Center Drive, San Diego, CA, USA.
Email: shtyllab@gmail.com

Present address
Tyler Cassidy, School of Mathematics, University of Leeds, Leeds, UK

Abstract

In drug development, quantitative systems pharmacology (QSP) models are becoming an increasingly important mathematical tool for understanding response variability and for generating predictions to inform development decisions. Virtual populations are essential for sampling uncertainty and potential variability in QSP model predictions, but many clinical efficacy endpoints can be difficult to capture with QSP models that typically rely on mechanistic biomarkers. In oncology, challenges are particularly significant when connecting tumor size with time-to-event endpoints like progression-free survival while also accounting for censoring due to consent withdrawal, loss in follow-up, or safety criteria. Here, we expand on our prior work and propose an extended virtual population selection algorithm that can jointly match tumor burden dynamics and progression-free survival times in the presence of censoring. We illustrate the core components of our algorithm through simulation and calibration of a signaling pathway model that was fitted to clinical data for a small molecule targeted inhibitor. This methodology provides an approach that can be tailored to other virtual population simulations aiming to match survival endpoints for solid-tumor clinical datasets.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE OF THIS TOPIC?

Algorithms have been established for exploring parameter uncertainty via the selection of virtual patients for quantitative systems pharmacology models.

WHAT QUESTION DID THIS STUDY ADDRESS?

This study investigates algorithmic modifications that are needed for existing virtual population algorithms to address specific complexities in the oncology drug development space.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

This study presents an improved approach for virtual patient selection for matching QSP model outputs to solid-tumor clinical data that accounts for complexities in linking tumor dynamics to time-to-event outcomes like progression-free survival.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT AND/OR THERAPEUTICS?

Improved virtual population selection algorithms within oncology drug development can help to better inform development decisions by improving efficacy projections, aiding clinical trial design via simulation, and identifying and interpreting causes of treatment resistance.

INTRODUCTION

Quantitative systems pharmacology (QSP) models have emerged as an important tool for interpreting and predicting the performance of oncology therapies.[1–4] Systems-level models help bridge the gap between our increasing knowledge of biological mechanisms—including oncogenic drivers, drug action, and resistance—and clinical response to therapy.[4,5] In oncology especially, understanding response variability due to resistance is essential for leveraging learnings from past trials and making accurate projections for new therapies or populations. Virtual populations (Vpops) have emerged as a useful tool for capturing patient heterogeneity with QSP models and are increasingly being used to inform drug development decisions.[1,2,6,7] Vpops consist of parameter samples (i.e., virtual patients) selected to match clinical measurement distributions while accounting for parameter and biological uncertainty.[1] The distribution of parameters selected within a Vpop can contain key biological insights, as each parameter has a biological interpretation and thus may be suggestive of important processes involved in treatment response or resistance. Predictions made using Vpops can also be used to quantify the expected variability for novel therapies, a valuable analysis for guiding development decisions. However, Vpop generation methods tailored to oncology models and end points are lacking, and further development in this area is needed.

Here, we focus on challenges in building Vpops for solid-tumor applications, where QSP model predictions of tumor size dynamics can be compared with patient sum of longest diameter (SLD) time series that are collected from sequential measurements of target lesions in many oncology trials. The observed SLD time series often change over months or years and interpreting them can be complex due to the multiple lesions and resistance processes involved.[8,9] Accurately capturing these time series with a virtual population can be especially difficult as they contain (i) a large degree of inter-patient variability in observed SLD trajectories, (ii) early truncation of trajectories for patients that progress due to target lesion tumor growth above clinically established RECIST SLD threshold, (iii) truncation of observed tumor trajectories due to non-SLD related progression causes, such as metastasis or non-target growth, or due to censoring.[10] This complexity poses challenges for selecting virtual populations as a model must be calibrated to capture very diverse tumor trajectories, but the comparator time series are structured such that there are fewer observations for more rapidly growing tumors. If not properly accounted for this sparsity, structure increases the risk of bias in the virtual population selection. Many of the truncations of the SLD time series are due to RECIST-defined progression events.[10] These progression events are important to predict as the primary efficacy end point in solid-tumor oncology trials is generally quantified as a time-to-event measurement, such as progression-free survival (PFS).[10] Predicting PFS as a QSP model output is therefore of great value for development decisions, but PFS events are only partially determined by SLD dynamics, and other events such as metastasis, censoring, or death also influence PFS. Careful consideration is therefore required to model the link between tumor size dynamics and primary efficacy end points like PFS.

In past publications, we have proposed Vpop generation approaches that employ probabilistic samplings of parameter space[1,2] and other groups have proposed methods based on prevalence weighting.[6,7] However, existing methods are not generally applicable to capture the statistical complexity of SLD and PFS end points in solid-tumor oncology, which involve interrelated time series and time-to-event end points subject to censoring. In this work, we propose an extension to the Rieger et al. Metropolis–Hasting (MH) algorithm that can be used to capture both aspects of the time-varying SLD dynamics and time-to-event end points like PFS with censoring.[2] In virtual trial simulation with large QSP models,[11] it is common to approach survival-type end points using some form of

approximation (i.e., equating the survival event with a specific tumor size threshold[12–14]), or focusing on response rates which are approximated from tumor size alone.[15–17] These approximations can serve as useful guides for many analysis goals, but they can pose challenges when attempting to precisely match clinical observations. In clinical datasets, censoring and metastasis events invariably occur, and there is an interdependence between the response rate and the times at which patients exit the trial. Accounting for these complexities is therefore important for efficacy projections to guide development decisions. Work in pharmacometrics using joint models of tumor burden time series and survival end points can account for some of these complexities and can help assess the correlation of covariates with tumor dynamics and survival end points.[18] However, further method development is needed as these methods are not readily applicable to typical large mechanistic QSP models.

While our methodology was designed to be broadly applicable to solid-tumor oncology QSP models, we will illustrate the steps involved in building a Vpop using a relatively simple mechanistic signaling and tumor growth model that predicts the time-varying tumor growth trajectory, illustrated in Figure 1. The QSP Vpop for this example was calibrated to data from a non-small cell lung cancer (NSCLC) phase II study of an ALK inhibitor, consisting of an SLD time series and censored PFS measurements.[19] We show that our Vpop selection algorithm can sample the parameters of the mechanistic model to capture patient heterogeneity in both SLD and PFS response and account for the dependence between the SLD trajectory and the PFS end point of primary interest using a probabilistic approach. We also demonstrate many of the advantages of this probabilistic approach, including making bootstrap predictions about PFS uncertainty, suggesting putative biomarkers, and informing future study planning.

METHODS

Dataset used for fitting

We used published data from a phase II study of an ALK inhibitor[19] in patients with NSCLC. The dataset consisted of SLD time series for each patient as well as a time-to-event measure indicating the time of patient exit from the trial and the cause, whether due to progression or censoring. We specifically focused our analysis on the subset of patients with at least two SLD measurements. Our dataset was also truncated at the 500th day of treatment, with all patients who remained in the trial up to this point being right-censored.

Signaling pathway and tumor growth inhibition

The QSP model used in this work consists of two core components: a minimal signaling MAPK/PI3K module and a tumor growth inhibition module where the cell

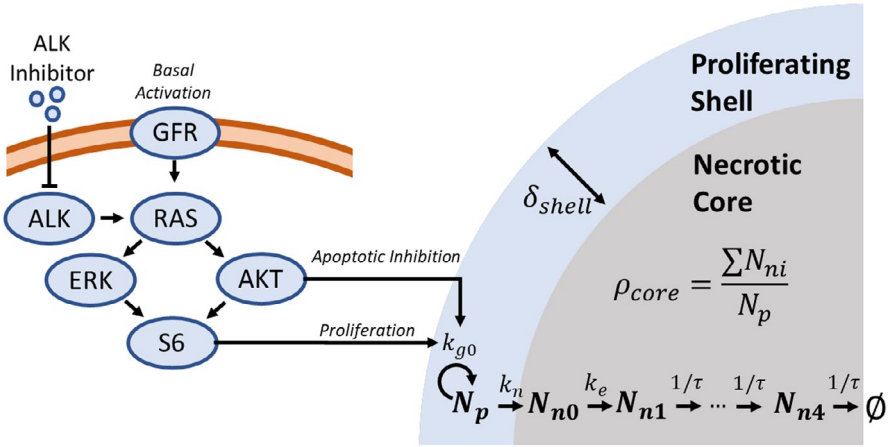

FIGURE 1 Structure of the mechanistic quantitative systems pharmacology (QSP) model used for virtual populations (Vpop) selection. Shown above is the cell signaling component, that captures the ALK inhibitor drug effect, and its connection to net cell growth/death in the proliferating shell compartment of a shell-core tumor growth model.[20] Here, transition to the necrotic compartment is presumed to be irreversible. The tumor physiology is controlled by the proliferating shell diameter, δ_{shell}, and the ratio of necrotic cells in the core to the proliferating cells in the shell, ρ_{core}. Here, N_p is the proliferating cell count, N_{ni} are the necrotic cell counts across the initial necrotic compartment and four subsequent clearance compartments, k_{g0} is the proliferation rate, k_n is the transition rate to the necrotic core, k_e is elimination rate into the clearance compartments, and τ is the timescale of necrotic clearance. The sum of longest diameter (SLD) for each simulated patient is defined as the diameter of the combined shell-and-core mass.

growth rates are controlled from signaling component concentrations. A graphical summary of the model is given in Figure 1 and further details are provided in the Supporting Information. In brief, the signaling model captures a minimal pathway of ALK downstream signal propagation through simplified RAS, ERK, and AKT components that drive the proliferation and apoptosis rates of cancer cells. A generic growth-factor receptor (GFR) term captures receptor tyrosine kinase (RTK) upstream signaling that drives the MAPK pathway through RAS/ERK. The signaling model, in turn, drives the shell-and-core tumor growth model via the net proliferation/death rate of cells in the tumor shell. The shell-and-core model builds on past works using a similar structure.[20–22] The SLD for each patient was modeled as the diameter of the combined shell-and-core cell volumes assuming a single spherical tumor shape per patient (Figure 1). A summary of the model and mathematical details can be found in Section S1. Parameters were included in the virtual population based on sensitivity analysis and consideration of known resistance mechanisms. The parameters selected to vary in the Vpop, along with their bounds, are shown in Table S1. Other parameters were held fixed at their nominal value which are given in Table S2.[23,24] The signaling model is driven by a simulated median steady-state drug plasma concentration–time course from a previously published model.[23] The plasma concentration directly modulates downstream signaling from ALK, see Section S1 for details.

Parameter sampling strategy to generate plausible patients

Our algorithmic approach is an extension of the two-step procedure described in Allen et al. and extended in Rieger et al.[1,2] The algorithm presented here consists of a plausible patient generation step, which uses a modified version of the MH implementation of Rieger et al.[2] Our base implementation of the MH algorithm is taken from Rieger et al. but novel additions have been made to sample time-to-event end points like PFS within each MH iteration. This approach adds a specialized scoring procedure for the PFS end point, one which respects RECIST criteria for progression based on primary tumor size but which also uses resampling of observed events and censoring labels to account for other risks. These novelties are highlighted in Figure 2 (right side). Plausible patients generated from the initial step conform to all prior bounds on parameters and model outputs and additionally have been selected for similarity in distribution to the observed clinical data via MH scoring. As in Rieger et al., a second stage is then used to further subsample plausible patients into a final virtual population via an acceptance–rejection algorithm to match the observed clinical distributions more precisely than the initial plausible population. Figure 2 provides a visual overview of our algorithm.

We compare candidate plausible patients against the observed population data on three measured outputs from the clinical dataset, namely baseline SLD, best percentage change in the SLD time series, and dropout time. Baseline SLD measurement is an SLD recorded prior to the start of therapy. Best percentage change is defined as the smallest observed on-trial percentage change in SLD from baseline for each patient. We define dropout time as the time to progression or censoring whichever was observed first, and these therefore correspond to the time patients are removed from the study for any cause. Together, best percentage change, which quantifies the change in tumor size, and dropout time, which quantifies the timing of the change, jointly provide information about the growth kinetics of the tumor. We have specifically focused on these outputs as best percentage change can be computed from available waterfall plots, and a dropout time distribution can be imputed from reported Kaplan–Meier (KM) curves—allowing for the potential use of published datasets alongside the individualized internal data used here.

To quantify the relative likelihood of various dropout time and best percentage change combinations for each plausible patient, a two-component Gaussian mixture was fit simultaneously to the three end points—baseline SLD, best percentage change, and dropout time—from the observed dataset. The fit Gaussian mixture distribution is shown as the blue contours in Figure 4. The baseline SLD was log-transformed prior to fitting as it is strictly positive and approximately log-normal. The dropout time is also strictly positive; however, better fits were achieved using the original scale; minimal probability mass (<2%) was allocated to non-positive dropout times. We experimented with a variety of distribution types along with Gaussian mixtures with a range of components and we observed the best fit and performance with a two-component mixture. However, the algorithm was sensitive to the type of distribution selected and may require careful tuning for new datasets. The fit mixture distribution was used to compute an acceptance score for each best percentage change and dropout time combination associated with a given plausible patient. The acceptance score is proportional to the probability density value of the fit Gaussian mixture distribution evaluated at the candidate baseline SLD, best percentage change, and dropout time values.

To compare plausible patients to the observed data, values for the above three outputs—baseline SLD, best percentage change, and dropout time—need to be computed from the model for each plausible patient. In our

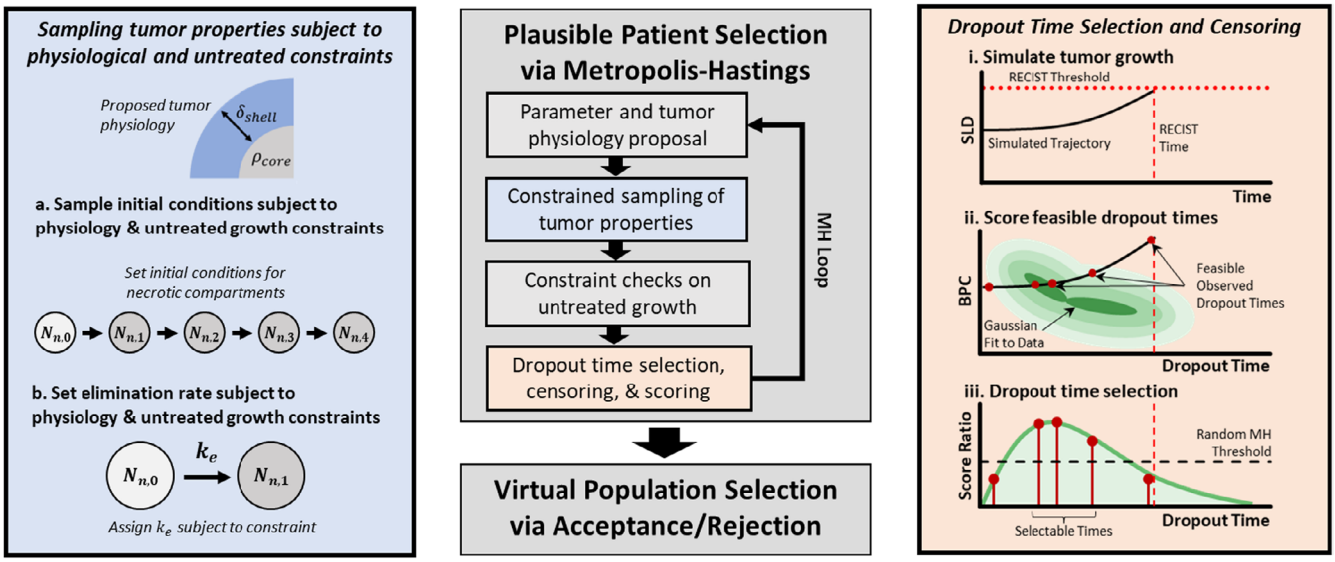

FIGURE 2 Overview of the virtual population selection procedure including (left) the procedure for sampling tumor physiology and enforcing untreated growth constraints, and (right) the procedure for selecting dropout times and assigning censoring labels. SLD, sum of longest diameter; BPC, best percentage change (in SLD); MH, Metropolis–Hasting.

algorithm, plausible patients are defined by their unique parameter values and tumor initial conditions. The plausible patient parameter sets are proposed in each MH iteration from a uniform proposal distribution centered at the last accepted patient and spanning 10% of the bounded parameter ranges in each sampled parameter dimension. Proposed parameter values are constrained to fall within the identified plausible ranges listed in Table S1. To capture non-signaling sources of variability in tumor physiology we also sample the core-shell ratio, ρ_{core}, and the shell thickness, δ_{shell}, in the Vpop directly, see Figure 1 and Section S2 for notational details. The values for ρ_{core} and δ_{shell} are therefore proposed alongside the model parameters in Table S1 at the beginning of the MH iteration, each with a uniform proposal distribution. Plausible ranges for ρ_{core} and δ_{shell} are listed in Table S3 and are enforced for each proposed pair of values. These physiological values also impose constraints on the initial conditions of the proliferative and necrotic cell populations. The initial conditions are randomized subject to these derived constraints and to ensure there is no untreated spontaneous tumor shrinkage. This procedure is depicted in the left box of Figure 2. Further details of the constrained randomization procedure can be found in Figure S1. After the initial conditions are set, the baseline SLD value for the plausible patient is calculated from ρ_{core} and δ_{shell}, see Section S2 for details. Once the parameter values and initial conditions have been proposed, subject to the above plausibility constraints, untreated growth is also simulated for 6 weeks to ensure a plausible doubling time in the absence of treatment. Plausible untreated doubling time ranges are listed in Table S3 and any candidate plausible patient that violates the untreated growth constraints, is rejected in the MH iteration.

Given the proposed parameter values and initial conditions for the plausible patient, a tumor trajectory under treatment is then simulated. This trajectory is truncated according to the RECIST criteria of 20% growth over nadir or the 500th treatment day, whichever occurs first. The computed RECIST threshold time or the 500th day represents the latest possible time a plausible patient could have an assigned dropout time. Earlier dropout times are possible or even likely depending on the tumor trajectory and the population being matched. For example, causes of progression under RECIST 1.1 criteria, include a >20% increase in SLD over nadir for target lesions (the change must also be >5 mm), discovery of a new lesion, qualitative assessment by the clinician of progression in non-target lesions, or death.[10] Importantly, the assignment of earlier dropout can influence the best percentage change end point for plausible patients, as earlier dropout times will be associated with relatively smaller changes in tumor size and therefore smaller best percentage change magnitudes.

To account for the risk of earlier dropout prior to reaching the RECIST threshold or the 500th day of treatment, for each plausible patient we sample from the distribution of observed dropout times from the actual study. For the given simulated parameter set in the algorithm, we consider truncating the tumor growth trajectory at any observed dropout time from the study occurring before the trajectory reaches the on-target RECIST threshold or 500th day. Therefore, if the RECIST threshold is reached, only observed times earlier than the point of threshold are

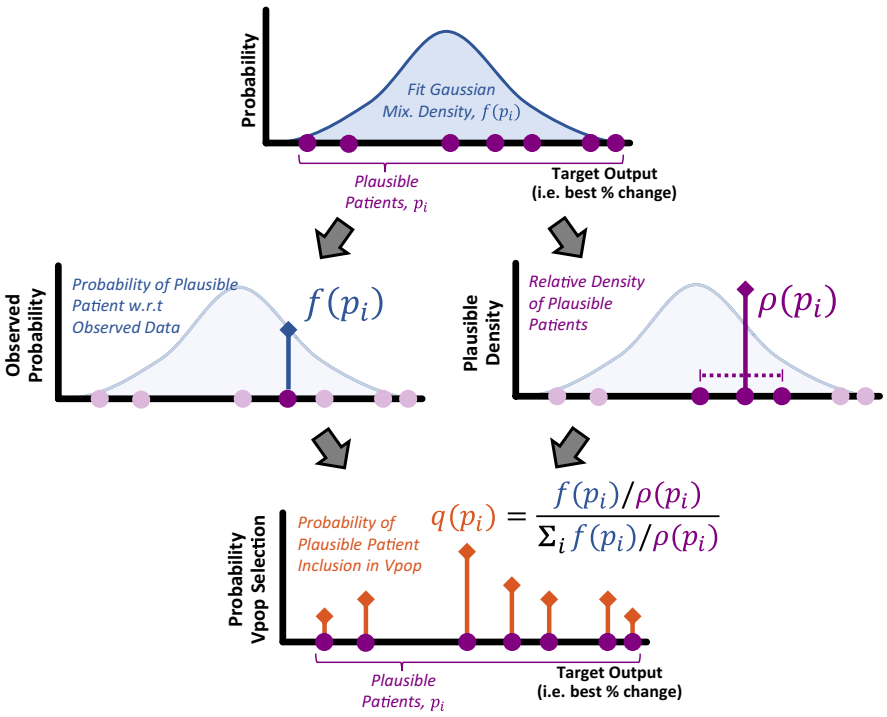

FIGURE 3 (A) A diagram illustrating our approach for computing the probability of plausible patient inclusion in a virtual population. Here, $\mathbf{p_i}$ is a vector of scored model outputs for the ith plausible patient, $f(.)$ is the Gaussian mixture density fit to the observed output data, $\rho(.)$, is local density estimate for plausible population in the neighborhood around the ith plausible patient, and $q(.)$ Is the probability of inclusion for the ith plausible patient. A full description of this approach is given in Section S3.

considered, otherwise, all observed times are considered. Each feasible dropout time implies a specific best percentage change for the simulated trajectory and plausible patient. At this point in each MH algorithm iteration, the potential plausible patient—with a fixed parameter set, initial conditions, and growth trajectory—has a range of feasible dropout times and best percentage change combinations, each varying in likelihood relative to the observed distribution from the clinical data. This procedure is summarized in the right box of Figure 2.

After scoring each feasible dropout time and best percentage change combination from the study for the given plausible patient trajectory, a single dropout time is selected probabilistically for the given plausible patient. This is done by first generating a uniform random threshold in the range from zero to one, similar to a standard MH algorithm, see Rieger et al. for details on MH used for virtual populations[2] and Chib et al. for general background on the MH algorithm.[25] The acceptance score ratio of each feasible dropout time and best percentage change for the current patient trajectory is then computed relative to the score for the previously accepted plausible patient from the prior MH iteration. A single dropout time is then randomly selected with equiprobability from the feasible set of dropout times that have acceptance score ratios above the random threshold. Dropout times with a score ratio below the random threshold are not considered for selection. If no dropout times have a score ratio meeting the random threshold, the plausible patient is rejected. For an accepted dropout time, the censoring status of the observed patient from which that accepted dropout time was sampled is also assigned to the plausible patient, as either censored or progressed. This allows the algorithm to account for censoring in the resulting virtual population. All together the proposed parameter set, initial conditions, dropout time, and censoring label are recorded together for the given accepted plausible patient.

Virtual population sampling strategy

As in prior work, a large set of plausible patients (here 10,000) are initially selected using the MH procedure outlined above.[2] This step provides a large collection of parametrically diverse plausible patients spanning the range of outputs observed in the clinical data. Accepted plausible patients are then subsampled using the acceptance–rejection algorithm described in Allen et al.[1] Scoring at this stage is only applied to the single assigned dropout time selected with each plausible patient in the MH algorithm. Readers are referred to the previous articles for a full description of the acceptance–rejection algorithm,[1,2] but in brief, the same Gaussian mixture fit to the observed data is used to acceptance score and subsample the overall plausible population into a final virtual population with an improved fit. The number of virtual patients selected by the Allen et al. and Rieger et al. acceptance–rejection algorithm is nondeterministic, but for the number of plausible patients considered here, it invariably exceeds the target sampled size for the dataset being matched.[1,2] We therefore use equal-probability down-sampling without replacement from the original virtual population to select a smaller virtual

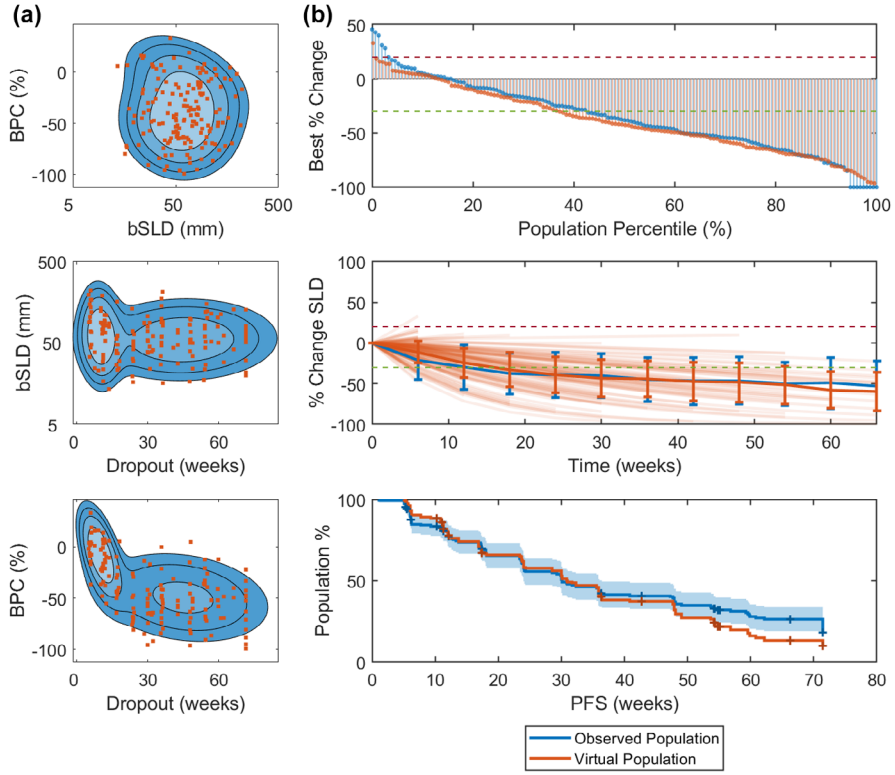

FIGURE 4 Example virtual population of 155 virtual patients, selected via our two-step algorithm, compared with observations from clinical data used for selection.[19] This example serves as a visual predictive check in comparing a realization of the virtual population with the observed population. (a) Left, the distribution of individual virtual patients is overlayed on the contours of the Gaussian fit to the clinical data used for scoring, showing good agreement in distribution and correlation. (b) The virtual and observed populations are compared on common visualizations for clinical data including a waterfall plot for best percentage change, median SLD time series with standard deviations, and Kaplan–Meier estimates for progression-free survival (PFS) functions.

population with the same sample size as the observed population when required in the results below.

A single virtual population of fixed sample size, as described above, can be useful for simulating novel therapies or trial designs for a fixed sample of virtual patients. However, by taking advantage of the probabilistic nature of our approach and repeatedly sampling virtual populations from the plausible population we can gain a better understanding of the expected variability of clinical end points over multiple simulated trials of a specified sample size and design. To efficiently enable this sampling, we use a simplified approach derived from Allen et al. to bootstrap-resample multiple, variable-sized, virtual populations from the plausible population. A graphical summary of our approach is shown in Figure 3 and further details are provided in the Section S3. In brief, the *probability of inclusion* for a plausible patient being included in a given virtual population is proportional to the likelihood of the plausible patient's clinical outputs (baseline SLD, best percentage change, and dropout time) with respect to the fit distribution to the observed data, and inversely proportional to the relative density of plausible patients in the local neighborhood of clinical outputs. The reader is referred to Allen et al. for the full derivation of the algorithm for computing the probability of inclusion but a summary of the procedure is provided for completeness in the Section S3.[1] The probability of inclusion can be used to resample the plausible population repeatedly allowing us to estimate how a given summary end point or statistical test will vary with sample size. We use this procedure to compute the expected uncertainty in the estimated median PFS for a given cohort size and perform a power analysis, as described in the Results. Our power analysis procedure involves simulating repeated virtual cohorts of varying sample sizes. On each of these cohorts, we perform the given statistical test we are interested in estimating the statistical power for, and over the set of virtual cohorts of a given size, we compute the fraction achieving statistical significance which yields an estimated power for a trial of the prescribed size.

RESULTS

To illustrate our algorithm, we first generated 10,000 plausible patients using our adapted MH approach. Generating

the set of plausible patients is the costliest component of the overall procedure, and the required number of plausible patients can vary depending on the parametric size of the mechanistic model; the relation between model complexity and the needed number of plausible patients is not generally known. However, the simulated set of plausible patients can be reused to select multiple virtual populations using the bootstrapping procedure described above. Here, we used diagnostics plots of mixing and convergence across the plausible sample, Section S4, to guide selection of the appropriate number of plausible patients. We found that 10,000 plausible patients allowed for dense and uniform coverage in the marginal pairwise parameter plots (Figure S2). Furthermore, we observed that autocorrelation for the MH chain parameter values decreased and then stabilized after several 100 plausible patients (Figure S3) and that, with 10,000 plausible patients, the individual parameter values over the course of the chain showed thorough mixing across each set of parameter bounds (Figure S4).

From the plausible population, we used the acceptance–rejection algorithm from Allen et al. to select a virtual population, which was then down-sampled to match the sample size of the target dataset. The acceptance–rejection step is needed because our MH proposal distribution is not necessarily symmetric, which is technically required for convergence to the target distribution given the acceptance threshold used here.[25] Enforcing a symmetric proposal for large QSP-type models is difficult which motivates the previously developed two-step procedure.[2] Here, we also down-sample to the target cohort size to aid in the interpretability of the visual predictive checks when comparing the simulated and observed cohorts. An example of a resulting virtual population is shown in Figure 4 alongside the clinical data used for selection. Figure 4a shows how the virtual population (orange points) captures the correlation structure between targeted clinical outputs from the observed data. The observed data distribution is shown here as the contours (blue) of the fit Gaussian mixture used for scoring. The virtual and observed populations also show good agreement in key clinical end points when compared on common visualization metrics including best percentage change shown in the waterfall plot, the median SLD time series, and PFS survival functions. The survival functions for both the observed and virtual populations were computed using a KM estimator to take account of the observed and simulated censoring. The KM estimators for the observed and virtual populations show good agreement in Figure 4b.

Multiple virtual population sampling

Using the Vpop resampling approach described in Methods, we compute a probability of inclusion for each plausible patient into a virtual population. We use these probabilities to efficiently sample multiple virtual populations to better quantify variability across virtual trials. For example, in Figure 5 we quantified the expected PFS end point variability, visualized as the 99th percentile interval for the survival probability over each week. This interval was computed for each week as the 0.5th and 99.5th percentiles of the weekly KM estimate of survival probability computed from 1000 bootstrapped virtual populations of 155 virtual patients each. These virtual populations were selected with replacement from the plausible population according to the probabilities of inclusion computed as in Figure 3. Also shown is the 99th percentile interval for the median survival time computed for the same set of simulations, which contains the observed median survival time and overlaps with the Greenwood 99% confidence interval computed from the observed data.

The probabilities of inclusion computed for our bootstrapping approach can also be used to explore, using the mechanistic signaling pathway model, potential mechanistic drivers of response within a clinical population of interest that our virtual trials match. These causes are not necessarily observable in the clinical population, as not all relevant biomarkers are measured, but in the virtual population, we are able to infer what specific combinations of model parameters drive predicted response. For example, we categorize virtual patients with a best percentage change $<-30\%$ as *responders* and patients with a best percentage change $>-30\%$ as *non-responders* and seek to understand the distributional differences

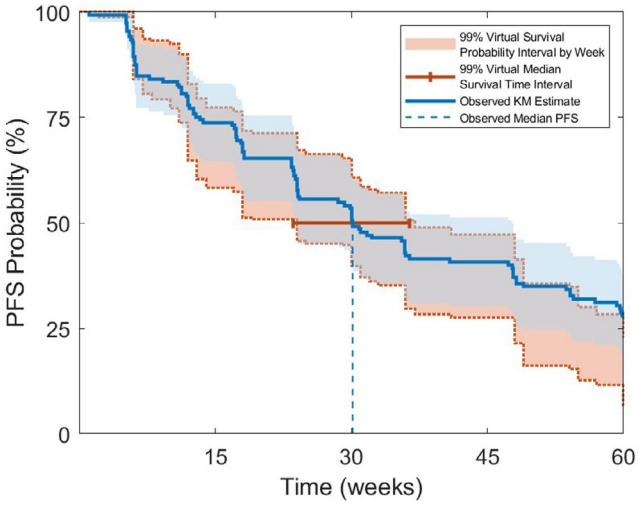

FIGURE 5 Plot of the Kaplan–Meier estimate and 99% confidence interval for the targeted clinical cohort in Solomon et al.[19] is shown in blue. Using our bootstrapping approach, the 99th percentile progression-free survival (PFS) interval computed by week for 1000 virtual trials of 155 virtual patients is shown in orange, along with the estimated 99th percentile interval for the median survival time shown for the same set of simulations.

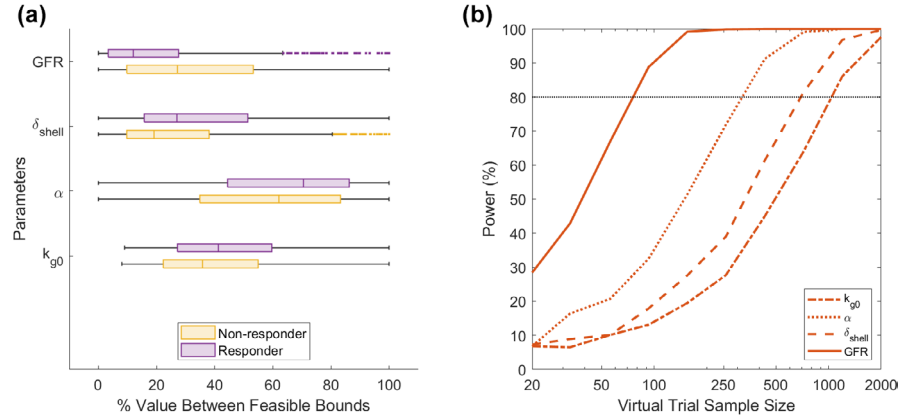

FIGURE 6 Analysis of variability in virtual population parameter values using bootstrap resampling of the virtual population (a) The distribution of the most significant parameters in differentiating responder (best percentage change $<-30\%$) and non-responder (best percentage change $>-30\%$) populations according to a Wilcoxon rank-sum test of the weight-resampled plausible population. (b) Power analysis showing the influence of sample size on power to detect significant differences in the top-ranked parameters from 6A in bootstrap simulated virtual trials of varying sample sizes.

between these two groups in terms of mechanistic parameter values. Comparing outcome subgroups on their parametric differences in a fixed-size virtual population (i.e., Figures 4 and 5 with $N=155$) can be highly variable due to the small number of patients in each comparison relative to the parameter dimensionality and end-point variability. To compute an estimate of the underlying subgroup difference in parameter values across the relevant clinical population at large, in the limit of large sample size, we selected a large, weighted resampling (10,000 patients with replacement) from the plausible population using the probabilities of inclusion as weights. We refer to this population as the *weight-resampled population*. This resampling procedure approximates the end point and parameter distributions of the smaller fixed sample size virtual populations but in the large sample size limit so that virtual trial-to-trial variability does not affect estimates of differences in parameters among population subgroups.

Using this approach, we performed a Wilcoxon rank-sum test between responder and non-responder patients in the weight-resampled plausible population and ranked parameters on their significance. Figure 6a depicts the four most significant parameters and visualizes the parameter quartiles for both the *responder* and *non-responder* populations. We see that the GFR is the most significant parameter in determining response; the GFR distribution demonstrates large variability in both groups, but there is an increase in median GFR activation among non-responders.

While Figure 6a shows potential drivers of differing virtual patient response, these differences may not necessarily be detectable in a realistic clinical population given the inherent parameter variability and bounds on realistic sample sizes. This may be relevant to planning biomarker measurements for future trials and informing trial size and design. To better understand the detectability of these inferred parameter differences between responder subgroups in realistic trials, we performed a power analysis to understand how large a sample would be needed to detect a significant difference in the top four parameters between the responder and non-responder populations using a Wilcoxon rank-sum test and a significance level of 5%. Figure 6b shows these results simulated using 1000 virtual trials with the specified sample size on the horizontal axis. The computed curves represent probabilities of correctly detecting a significant difference for the given sample size and illustrate that the difference in a composite measure of GFR activity could likely be detected with 80% power using fewer than 100 patients, whereas the differences in δ_{shell} α, and, k_{g0} would require larger trials of hundreds of patients to detect. While the exact nature of GFR activity measurement would add additional qualifications to this analysis, this approach can serve as an initial guide for what drivers of resistance may be important and detectable in future studies.

DISCUSSION

Virtual populations are becoming an increasingly important development tool for analysis and prediction using QSP models. In this work, we propose an extended virtual population selection procedure for solid-tumor oncology trials. Our procedure ensures constrained yet interpretable sampling of tumor physiology and provides a novel, data-driven approach to account for both censoring and non-target progression risks when matching PFS end points.

Our selection procedure can match important measures of efficacy in oncology including best percentage change (i.e., waterfall plots) and PFS (i.e., survival curves) using the demonstration dataset. We further use the probabilistic sampling approach inherent in our method to demonstrate how the plausible and virtual populations can be leveraged to understand patterns in parameter values between responder subgroups and across trials of varying size. At this time, we advise that care is needed if using this approach with new therapeutic modalities or indications. Careful checking of virtual population agreement with any targeted clinical data is important when applying the method to new datasets, and further validation would be ideal before the method is used for prospective extrapolation across modalities or indications. Also, in this work we do not incorporate pharmacokinetic variability, instead using a median model. In future work, we aim to benchmark our approach in terms of predictive accuracy and computational efficiency to make it more general and broaden its applicability.

AUTHOR CONTRIBUTIONS

N.B., B.S., T.J., T.C., R.K., M.T., R.A., and C.J.M. wrote the manuscript. B.S., T.C., N.B., R.A., and M.T. designed the research. N.B. and T.J. performed the research. K.P., T.J., and N.B. analyzed the data.

FUNDING INFORMATION

This work was funded by Pfizer Inc.

CONFLICT OF INTEREST STATEMENT

N.B., M.T., K.P., R.A., C.J.M., and B.S. are employees of Pfizer Inc., and T.C. was an employee of Pfizer Inc. during this work. R.K. is an employee of Vantage Research Inc., and T.J. is affiliated with Vantage Research Inc. in the capacity of an independent contractor.

ORCID

Tanvi Joshi https://orcid.org/0000-0002-9729-8979
Kamrine Poels https://orcid.org/0000-0002-7220-2213
Cynthia J. Musante https://orcid.org/0000-0002-3003-0169

REFERENCES

1. Allen RJ, Rieger TR, Musante CJ. Efficient generation and selection of virtual populations in quantitative systems pharmacology models. *CPT Pharmacometrics Syst Pharmacol*. 2016;5(3):140-146.
2. Rieger TR, Allen RJ, Bystricky L, et al. Improving the generation and selection of virtual populations in quantitative systems pharmacology models. *Prog Biophys Mol Biol*. 2018;139:15-22.
3. Azer K, Kaddi CD, Barrett JS, et al. History and future perspectives on the discipline of quantitative systems pharmacology modeling and its applications. *Front Physiol*. 2021;12:637999.
4. Chelliah V, Lazarou G, Bhatnagar S, et al. Quantitative systems pharmacology approaches for Immuno-oncology: adding virtual patients to the development paradigm. *Clin Pharmacol Ther*. 2021;109(3):605-618.
5. Kirouac DC, Schaefer G, Chan J, et al. Clinical responses to ERK inhibition in BRAF(V600E)-mutant colorectal cancer predicted using a computational model. *NPJ Syst Biol Appl*. 2017;3:14.
6. Gadkar K, Budha N, Baruch A, Davis JD, Fielder P, Ramanujan S. A mechanistic systems pharmacology model for prediction of LDL cholesterol lowering by PCSK9 antagonism in human Dyslipidemic populations. *CPT-Pharmacometrics Syst Pharmacol*. 2014;3(11):1-9.
7. Schmidt BJ, Casey FP, Paterson T, Chan JR. Alternate virtual populations elucidate the type I interferon signature predictive of the response to rituximab in rheumatoid arthritis. *BMC Bioinformatics*. 2013;14:1–16.
8. Kumar R, Thiagarajan K, Jagannathan L, et al. Beyond the single average tumor: understanding IO combinations using a clinical QSP model that incorporates heterogeneity in patient response. *CPT Pharmacometrics Syst Pharmacol*. 2021;10(7):684-695.
9. Kumar R, Qi T, Cao Y, Topp B. Incorporating lesion-to-lesion heterogeneity into early oncology decision making. *Front Immunol*. 2023;14:1173546.
10. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228-247.
11. Craig M, Gevertz JL, Kareva I, Wilkie KP. A practical guide for the generation of model-based virtual clinical trials. *Frontiers in Systems Biology*. 2023;3:1174647.
12. Cassidy T, Craig M. Determinants of combination GM-CSF immunotherapy and oncolytic virotherapy success identified through treatment personalization. *PLoS Comput Biol*. 2019;15(11):e1007495.
13. Jenner AL, Cassidy T, Belaid K, Bourgeois-Daigneault MC, Craig M. In silico trials predict that combination strategies for enhancing vesicular stomatitis oncolytic virus are determined by tumor aggressivity. *J Immunother Cancer*. 2021;9(2):e001387.
14. Cardinal O, Burlot C, Fu Y, et al. Establishing combination PAC-1 and TRAIL regimens for treating ovarian cancer based on patient-specific pharmacokinetic profiles using in silico clinical trials. *Computat Syst Oncol*. 2022;2(2):e1035.
15. Wang HW et al. Conducting a virtual clinical trial in HER2-negative breast cancer using a quantitative systems pharmacology model with an epigenetic modulator and immune checkpoint inhibitors. *Front Bioeng Biotechnol*. 2020;8:141.
16. Wang H, Arulraj T, Kimko H, Popel AS. Generating immunogenomic data-guided virtual patients using a QSP model to predict response of advanced NSCLC to PD-L1 inhibition. *NPJ Precision Oncol*. 2023;7(1):55.
17. Arulraj T, Wang H, Emens LA, Santa-Maria CA, Popel AS. A transcriptome-informed QSP model of metastatic triple-negative breast cancer identifies predictive biomarkers for PD-1 inhibition. *Sci Adv*. 2023;9(26):eadg0289.
18. Ibrahim JG, Chu HT, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncol*. 2010;28(16):2796-2801.
19. Solomon BJ, Besse B, Bauer TM, et al. Lorlatinib in patients with ALK-positive non-small-cell lung cancer: results from a global phase 2 study. *Lancet Oncol*. 2018;19(12):1654-1667.

20. Checkley S, MacCallum L, Yates J, et al. Bridging the gap between in vitro and in vivo: dose and schedule predictions for the ATR inhibitor AZD6738. *Sci Rep*. 2015;5:13545.
21. Evans ND, Dimelow RJ, Yates JWT. Modelling of tumour growth and cytotoxic effect of docetaxel in xenografts. *Comput Methods Prog Biomed*. 2014;114(3):E3-E13.
22. Drexler DA, Sápi J, Kovács L. Modeling of tumor growth incorporating the effects of necrosis and the effect of bevacizumab. *Complexity*. 2017;2017:1-10.
23. Chen J, Houk B, Pithavala YK, Ruiz-Garcia A. Population pharmacokinetic model with time-varying clearance for lorlatinib using pooled data from patients with non-small cell lung cancer and healthy participants. *CPT Pharmacometrics Syst Pharmacol*. 2021;10(2):148-160.
24. Yamazaki S, Lam JL, Zou HY, Wang H, Smeal T, Vicini P. Translational pharmacokinetic-pharmacodynamic modeling for an orally available novel inhibitor of anaplastic lymphoma kinase and c-Ros oncogene 1. *J Pharmacol Exp Ther*. 2014;351(1):67-76.
25. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. *Am Stat*. 1995;49(4):327-335.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Braniff N, Joshi T, Cassidy T, et al. An integrated quantitative systems pharmacology virtual population approach for calibration with oncology efficacy endpoints. *CPT Pharmacometrics Syst Pharmacol*. 2025;14:268-278. doi:10.1002/psp4.13270