Systems biology

Complete populations of virtual patients for in silico clinical trials

S. Sinisi\textsuperscript{1,1}, V. Alimguzhin\textsuperscript{1}, T. Mancini \textsuperscript{1,*}, E. Tronci\textsuperscript{1}, B. Leeners\textsuperscript{2}

\textsuperscript{1}Computer Science Department, Sapienza University of Rome, Italy
\textsuperscript{2}Division of Reproductive Endocrinology, University Hospital Zurich, Switzerland

\* To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

26 November 2020

Abstract

Motivation. Model-based approaches to safety and efficacy assessment of pharmacological drugs, treatment strategies, or medical devices (In Silico Clinical Trials, ISCT) aim to decrease time and cost for the needed experimentations, reduce animal and human testing, and enable precision medicine. Unfortunately, in presence of non-identifiable models (e.g., reaction networks), parameter estimation is not enough to generate complete populations of Virtual Patients (VPs), i.e., populations guaranteed to show the entire spectrum of model behaviours (phenotypes), thus ensuring representativeness of the trial.

Results. We present methods and software based on global search driven by statistical model checking that, starting from a (non-identifiable) quantitative model of the human physiology (plus drugs PK/PD) and suitable biological and medical knowledge elicited from experts, compute a population of VPs whose behaviours are representative of the whole spectrum of phenotypes entailed by the model (completeness) and pairwise distinguishable according to user-provided criteria. This enables full granularity control on the size of the population to employ in an ISCT, guaranteeing representativeness while avoiding over-representation of behaviours.

We proved the effectiveness of our algorithm on a non-identifiable ODE-based model of the female Hypothalamic-Pituitary-Gonadal axis, by generating a population of 4 830 264 VPs stratified into 7 levels (at different granularity of behaviours), and assessed its representativeness against 86 retrospective health records from Pfizer, Hannover Medical School and University Hospital of Lausanne. The datasets are respectively covered by our VPs within Average Normalised Mean Absolute Error of 15\%, 20\%, and 35\% (90\% of the latter dataset is covered within 20\% error).

1 Background

Model-based approaches to safety and efficacy assessment of drugs, pharmacological treatments, or medical devices (In Silico Clinical Trials, ISCT) hold the promise to decrease time and cost for the needed experimentations, reduce the need for animal and human testing, and enable precision medicine, where personalised treatments or devices optimised for each patient can be designed before being actually administered or implanted (Avicenna Project, 2016; Puppulardo et al., 2019). To enable ISCT, quantitative mechanistic models (Virtual Physiological Human, VPH, models) of the human (patho-) physiology as well as of the relevant medicinal drugs are being actively developed and validated. Such models define drug concentration time courses and effects (Pharmacokinetics/Pharmacodynamics, PK/PD) and the physiology of interest at different levels of scale, ranging from molecules (e.g., Roy and Roy, 2010), molecular and gene networks (e.g., Le Novère, 2015), cells (e.g., Bachler et al., 2014), organs (e.g., Cox et al., 2009), up to body compartments (e.g., Balazki et al., 2018) and the whole body (e.g., Hester et al., 2011).

1.1 Motivation

One of the main enablers to perform an ISCT is the availability of a finite population of virtual patients, i.e., computational models able to predict (via simulation) relevant clinical measurements (those needed to assess efficacy/safety of the therapy, i.e., drug, treatment, or device, under trial) from time courses of clinical actions (such as drug administrations, see, e.g., FDA, 2018; EMA, 2019). For an ISCT to provide compelling evidence of the safety/efficacy of a therapy and to support its design and revision, such population must be complete, i.e., representative of the...
entire spectrum of behaviours deemed of interest, from both physiology and drug PK/PD points of view. Virtual Patients (VPs) are typically derived by parameterising quantitative mechanistic VPH models, which in turn are defined by encoding qualitative knowledge of the human physiology of interest (e.g., from the literature or pathway databases like KEGG, Kanehisa et al., 2017 or Reactome, Fabregat et al., 2018) as well as PK/PD of pharmaceutical compounds (e.g., Lippert et al., 2019) into mathematical systems such as, e.g., Ordinary Differential Equations (ODEs) or difference equations (see, e.g., Bartocci and Liò, 2016; Iruzun-Aran et al., 2017). Indeed, it is by means of parameters (such as stoichiometric constants, rates, or other patient-specific quantities) that such models take into account inter-subject variabilities, as different parameter assignments yield different model trajectories, also in terms of reactions to drug administrations.

1.2 State of the art in computing populations of VPs

Different approaches have been proposed to compute a population of VPs for quantitative VPH models. Such approaches greatly differ depending on whether the given model is identifiable or non-identifiable.

For identifiable models, a complete population of VPs can be computed by fitting the models against a set of in vivo measurements deemed representative of models. We focus on, for example, the Physiologically-based Pharmacokinetics (PBPK) simulator in (Lippert et al., 2019) provides a large set of VPs compliant with PBPK regulations from EMA, FDA, EFSA, or EPA. Also, in (Kovatchev et al., 2009) a VPH model is described, and a population of 300 VPs is provided for it, representing 100 adults, 100 adolescents, and 100 children. Such VPs have been approved by FDA as a substitute for pre-clinical animal testing of new treatment strategies for Type 1 Diabetes Mellitus. The above models enjoy a very important property: all their parameters describe physiological characteristics, have known ranges of values, and can be reliably estimated through in-vivo or in-vitro measurements.

The situation becomes more intricate for non-identifiable models, for which, to our knowledge, no approach is available to compute complete populations of VPs. In fact, although for such models parameter estimation can still be used (e.g., Teunoncos et al., 2015; Allen et al., 2016; Rieger et al., 2018; Schmaeuser et al., 2019; Wang et al., 2020) and citations thereof) to find cases (counterexamples) where the therapy under assessment is unsafe/ineffective, the resulting population of VPs is not guaranteed to be complete, no matter how large or representative is the input dataset used for fitting. This is because, due to model non-identifiability, there could be other (possibly very different) parameter assignments (not selected through fitting) still matching experimental data, but leading to different model behaviours under the new therapy.

In other words, model non-identifiability hinders the possibility to have a comprehensive picture of the cases where the therapy succeeds or fails. As a result, although being based on solid scientific principles (e.g., biochemical reactions), thereby satisfying one of the qualification requirements for ISCT (e.g., FDA, 2018; EMA, 2019), it is hard to use non-identifiable models to verify safety/efficacy of a therapy. This is why identifiability is a key test in, e.g., FDA or EMA PBPK guidelines.

In the literature, qualitative VPH models have also been considered, for example logic-based models (e.g., Wang et al., 2012; Bloomingdale et al., 2018). Their aim is to predict sequences of Boolean-valued (low vs. high) expression levels rather than the time course of the biological quantities of interest. In qualitative models, non-identifiability can somewhat be overcome by modelling lack of knowledge about reaction rates through an asynchronous update schema for their Boolean-valued variables. Complete populations of VPs can then be generated by using finite state model checking techniques to look for attractors (e.g., Zheng et al., 2013; Khan et al., 2017; Razzaghi et al., 2018 and citations thereof). Unfortunately, this approach cannot be used for quantitative models (like those defined through ODEs or difference equations, our main focus here) defining real-valued (rather than Boolean-valued) concentrations of compounds, where, in general the state space is infinite.

We finally argue that the above problem stemming from non-identifiability also arises in other areas. For example, models used in machine learning (e.g., neural networks) are typically non-identifiable, and it is well known that, notwithstanding how large is the training dataset, it is possible to find (plausible) input data leading to wrong classifications (e.g., Eykholt et al., 2018). Not surprisingly, similarly to ISCT, this is the main obstacle in qualifying machine learning–based approaches within safety-critical (i.e., high impact regulatory purpose) applications such as autonomous driving (e.g., Ienn et al., 2020).

The above considerations motivate the main goal of this paper: to develop methods and software that (possibly building on parameter estimation against in vivo data) can compute a finite set of physiologically meaningful, pairwise distinguishable VPs, which are representative of the entire spectrum of behaviours defined by the given (possibly non-identifiable) quantitative VPH model (completeness).

1.3 Contributions

In this paper we present methods and software to compute populations of VPs for (possibly non-identifiable) quantitative VPH models. We focus on the typical case of models that, due to their complexity, cannot be analysed symbolically, but need to be numerically simulated (e.g., Hucka et al., 2003; Maggioni et al., 2020), and show the effectiveness of our methods on a non-identifiable model of the Hypothalamic-Pituitary-Gonadal (HPG) axis defined in terms of 33 highly non-linear stiff ODEs.

Our populations satisfy three important properties: completeness, pairwise distinguishability, and stratifiedness.

Completeness means that our populations show all model behaviours deemed of interest (e.g., physiologically meaningful), even when such a full set of behaviours is unknown at model design time (this is typical in large non-identifiable, over-parametrisised VPH models, see below). For example, the population we computed in our case study comprises as many as 48,302,64 VPs.

Pairwise distinguishability means that no model behaviour (aka phenotype) is over-represented in our population: any two VPs behave differently (according to some used-defined notions of behavioural distinguishability) in at least one scenario (e.g., input pattern) supported by the model. This avoids waste of computation during an ISCT.

Stratifiedness means that our populations are organised in levels, (strata), each one showing the entire spectrum of behaviours under different distinguishability criteria. For example, in our case study we stratified our 48,302,64 VPs into 7 sub-populations, each one comprising a number of VPs ranging from 2 million to just 1. Since each sub-population alone is representative of the entire spectrum of model behaviours (of course at different granularity), proper trade-offs can be sought, when designing an ISCT, between the needed behavioural granularity and the budgeted computational effort.

Our any-time algorithm, based on global search guided by statistical model checking, intelligently explores the (typically huge) model parameter space, collects those parameter assignments showing a physiologically meaningful behaviour (i.e., VPs), and organises them into strata, while guaranteeing a statistically-sound form of graceful degradation.

Note that, in many non-identifiable models (like our case-study HPG axis model), most parameter assignments might not actually represent VPs, as, upon simulation, their associated model trajectories show-up to be physiologically meaningless or, anyway, out of interest. This is due to, e.g., over-parametrisation, presence of parameters whose values are not measurable through clinical assays (e.g., reaction rates), presence of unknown (hence, not modelled) interdependency constraints...
among parameters, and use of parameters to define not-well-understood physiological mechanisms. To find parameter assignments yielding physiologically meaningful model behaviours and different phenotypes is thus computationally very hard, and naive exploration or sampling of the parameter space could be hopeless.

In order to automatically recognise physiologically meaningful model behaviours (and thus parameter assignments defining VPs), our approach envisions the elicitation and formalisation of background biological and medical knowledge (possibly also coming from available data). Our approach is fully independent of how such knowledge is formalised, as long as we can define a criterion that, given a parameter assignment (a candidate VP), decides whether the resulting model trajectory is physiologically meaningful or not.

In our case study, we rely on background knowledge available in terms of known assignments to the system observables for the system output might depend on. Such criteria are applicable to a wide class of models, e.g., those defining hormonal signalling networks.

2 Material and methods

Below we define our framework (Section 2.1) and methodology (Section 2.2) to generate complete stratified populations of pairwise distinguishable VPs.

2.1 Formal framework

VPH models. We adopt a very general approach to define VPH models and view them as parametric input-output dynamical systems. This general definition is standard in signals and systems (see, e.g., Sontag, 1998), especially when, as in the case of physiological models, the system internal state is not accessible, and only selected outputs (system observables) can be measured.

Our definition (for a formal statement see Definition 1 in Section SM1.1.1 of the supplementary material) accounts for both continuous- and discrete-time models (e.g., those defined by means of ODEs and difference equations, respectively). Namely, model inputs are time functions \( u(t) \) defining the time course of exogenous inputs (e.g., drug administrations). Our models are parametric, in that their observation function \( y(u, \lambda) \), defining the values \( y(t); u, \lambda \) of the system observables at any time point \( t \), depends on both the input time function \( u \) and the values \( \lambda \) for the system parameters, chosen within the model parameter space \( \Lambda \).

For physical reasons, we require that our VPH models are strictly causal, i.e., their observation function up to any time point depends only on past inputs. Also, given the presence of parameters, we focus on deterministic systems, in that parameters embody any initial condition which the system output might depend on.

Virtual patients, phenotypes, populations. As anticipated in Section 1, not all assignments to a VPH model parameters yield behaviours of interest. Many might even yield physiologically meaningless behaviours. Conversely, due to, e.g., system over-parametrisation or non-identifiability, multiple parameter assignments may yield (almost) indistinguishable behaviours (i.e., their associated observation functions are very similar on all inputs). Such indistinguishable VPs would increase the computational effort needed to carry out an ISCT on the entire population, without bringing any advantage in terms of representativeness of the trial.

For generality, our forthcoming definitions rely on user-provided Boolean function \( \varphi \) and equivalence relation \( \sim \). Boolean function \( \varphi \) defines the conditions to be met by any parameter \( \lambda \in \Lambda \) for the associated model behaviours to be considered of interest, for example physiologically meaningful (in which case, \( \lambda \) has to be regarded as a VP). Equivalence relation \( \sim \) on the set of VPs defines when two VPs shall be considered having indistinguishable behaviour (i.e., showing the same phenotype): for any two VPs \( \lambda, \lambda', \lambda \sim \lambda' \) means that the two VPs show the same phenotype.

With respect to given \( \varphi \) and \( \sim \) for a VPH model \( S \), we define the following concepts (for a formal statement see Definition 2 in Section SM1.1.2): (a) the population \( \hat{\Lambda} \) of VPs for \( S \) is the set of parameter assignments \( \lambda \in \Lambda \) for which \( \varphi(\lambda) \) is true; (b) the phenotype of VP \( \lambda \) is the equivalence class of \( \lambda \) with respect to \( \sim \) (notation: \( [\lambda]_\sim \)); (c) the phenotype space \( \hat{\Lambda} \) of \( \hat{\Lambda} \) is the quotient set of \( \Lambda \) with respect to \( \sim \), i.e., the set of all-different phenotypes of VPs in \( \hat{\Lambda} \); (d) an All-Different Phenotype Population (APP) of VPs is any subset \( \hat{\Lambda}^* \) of \( \hat{\Lambda} \) such that no two VPs \( \lambda, \lambda' \) exist in \( \hat{\Lambda}^* \) having the same phenotype. Also, an APP \( \hat{\Lambda}^* \) is said a Complete APP (CAPP) if it contains a representative of all phenotypes in the phenotype space of \( \hat{\Lambda} \).

Clearly, the definition of both function \( \varphi \) and relation \( \sim \) depends on the VPH model at hand, and has to be made starting from expert knowledge. Also, in the typical case of models subject to external inputs (e.g., drug administrations), both \( \varphi \) and \( \sim \) might need to be defined on model behaviours under different input functions. This allows the expert to define meaningfulness and phenotypes of candidate VPs also in terms of their reactions under different sequences of drug administrations (where such reactions are dictated by the PK/PD model equations).

Note that, when \( \sim \) is \( = \) (i.e., the equivalence relation defining a distinct class per VP \( \lambda \in \Lambda \)), we have \( \lambda, [\lambda]_\sim = \hat{\lambda} \). Hence, the entire population of VPs \( \hat{\Lambda} \) can always be regarded as a CAPP.

In Section 3 we give a widely-applicable definition for \( \varphi \) and \( \sim \) based on qualitative similarity of the model evolutions associated to different parameters.

2.2 Computing complete populations of VPs

Given a VPH model with parameter space \( \Lambda \), a Boolean function \( \varphi \) and an equivalence relation \( \sim \) as in Section 2.1, our goal is to compute a CAPP with respect to \( \varphi \) and \( \sim \).

In this paper we focus on cases where the definition of the VPH model, function \( \varphi \), and the computation of the phenotype \( [\lambda]_\sim \) of a VP \( \lambda \) are too complex for set \( \hat{\Lambda}^* \) to be computed analytically and/or symbolically in closed form. For such complex scenarios, deciding whether \( \varphi(\lambda) = \text{true} \) or not for any given \( \lambda \in \Lambda \) (hence, whether \( \lambda \) represents a VP or not) and, in the affirmative case, computing its phenotype \( [\lambda]_\sim \), involves a numerical simulation of the VPH model and the subsequent analysis of the resulting model trajectories under different inputs. Also, knowing that \( \varphi(\lambda) = \text{true} \) for some \( \lambda \in \Lambda \) does not allow us to infer (without additional simulations) whether \( \varphi(\hat{\lambda}) = \text{true} \) for other parameters \( \hat{\lambda} \in \hat{\Lambda} \), let alone their phenotypes.

In order to cope with such a general setting, we adopt a search-based approach that explores the model parameter space \( \Lambda \) looking for parameters \( \lambda \in \Lambda \) such that \( \varphi(\lambda) = \text{true} \) and belonging to all-different equivalence classes of \( \sim \). This calls for VPH models whose parameter space \( \Lambda \) is finite or can be finitised by the user, e.g., into a bounded interval of \( \mathbb{R}^k, k > 0 \). Such finitisation can often be performed by exploiting knowledge about, e.g., physiological bounds to the parameter values and model locality assumptions (i.e., minor changes to the value of a parameter yield minor changes in the resulting model behaviours).

Nevertheless, even when \( \Lambda \) is finite, an exhaustive exploration is practically infeasible unless \( \Lambda \) is very small. Unfortunately, this is not the case for complex VPH models: for example, the size of the (finitised) parameter space of our case-study model is \( 10^{76} \), which makes an exhaustive search clearly out of reach (let alone the fact that computing \( \varphi(\lambda) \) for each \( \lambda \) takes seconds of simulation time).
To overcome these obstacles, our search (Section 2.2.1) is an any-time algorithm relying on Statistical Model Checking (SMC) and hypothesis testing to guarantee proper statistically-sound graceful degradation.

2.2.1 The algorithm

Our algorithm is an any-time procedure which builds on the SMC and hypothesis testing methods initially presented in (Grosu and Smolka, 2005) and extended in (Tronci et al., 2014).

Core algorithm. Given a VPH model $S$ having finite (although too large for an exhaustive exploration) parameter space $\Lambda$, plus function $\varphi$ and equivalence relation $\sim$, our algorithm implements a one-sided error procedure to compute a CAPP $\hat{\Lambda}$ for $S$ with respect to $\sim$. The algorithm randomly samples the parameter space $\Lambda$ (according to a user-defined sampling policy to the exploration of each slice, in order to keep the values of $\varepsilon$ those parameters $\lambda$ that represent VPs (i.e. $\varphi(\lambda) = \text{true}$) and show a phenotype different than all those already represented in $\hat{\Lambda}$.

The algorithm can be interrupted at any time and provides a form of graceful degradation: after each sample, the algorithm computes an upper bound $c \in [0,1]$ to the probability that further sampling would produce VPs of unseen phenotypes (error margin). This fact would prove that the current APP is not indeed a CAPP. When the achieved value for $c$ reaches a sufficiently-small (target) threshold, the user can decide to stop the algorithm and get the APP computed so far.

The computed value for $c$ is a function of the number of consecutive failed attempts $N$ that the algorithm is experiencing in discovering VPs of new phenotypes. Clearly, being based on sampling, our algorithm can commit an error in computing the error margin $c$ (i.e., it could return a value lower than a true upper bound). However, by exploiting statistical hypothesis testing methods, given any user-requested value $\delta \in (0,1)$ (confidence ratio), our algorithm ensures (see below and Theorem 1 in Section SM1.2.1) that the probability of such an error is at most $\delta$.

Sampling policy. In order to be effective in discovering VPs of new phenotypes, the employed sampling policy may embody proper domain expert knowledge and structural knowledge about the VPH model, for example: independence constraints among components of the parameter values (very common in over-parameterised models), or sensitivity information of model behaviours with respect to parameter values. Also, the sampling policy can be refined and improved during the search to embed new knowledge, e.g., about the newly discovered VPs. In Section 3.4 we will outline a sampling policy for our case-study model (but widely applicable in general), which exploits the above flexibility.

Parallel computation. Our algorithm takes advantage of a parallel High Performance Computing (HPC) infrastructure. The parameter space $\Lambda$ is split upfront into $k$ slices $\Lambda_1, \ldots, \Lambda_k$, and $k$ independent instances of our core algorithm can be run in parallel, where instance $i$ with $i \in [1,k]$ draws samples from $\Lambda_i$ to build population $\hat{\Lambda}_i$. When $\hat{\Lambda}_i$ is computed for all slices, a final population $\hat{\Lambda}$ is produced by taking the union of the phenotype spaces of all $\hat{\Lambda}_i$ and by choosing one representative VP from each equivalence class. To take load balancing into account, the overall number of parallel processes can be much higher than the number of slices $k$. An orchestrator can then dynamically assign such processes, in order to keep the values of $\varepsilon$ balanced. This approach to parallelism and load balancing is very effective (see, e.g., Manzini et al., 2016) and avoids overhead due to inter-process communication (as that experienced in, e.g., Manzini et al., 2015).

Simultaneous computation of stratified APPs. Our algorithm can work with multiple equivalence relations $\sim_1, \ldots, \sim_L$, defining different behavioural indistinguishability (i.e., same phenotype) criteria, e.g., at different levels of abstraction. When it makes sense to use the same policy to sample the VPH model parameter space $\Lambda$ for all the $\sim_l$ ($l \in \{1, L\}$), then the $L$ CAPPs can be computed simultaneously using the same sequence of random samples. In Section 3 we will exploit this possibility to compute a hierarchy of stratified CAPPs for our case-study VPH model.

Complete algorithm and main result. Let $\Lambda_1, \ldots, \Lambda_k$ be a partitioning of the finite (or finitely) parameter space $\Lambda$ of our VPH model $S$ into $k > 0$ slices. Our overall algorithm runs in parallel in instances of the algorithm in Figure 1, where instance $i \in [1,k]$ runs on slice $\Lambda_i$ of $\Lambda$ and computes $\hat{\Lambda}_i > 0$ APPs, one for each given equivalence relation $\sim_l$ ($l \in \{1, L\}$) on the population of VPs entailed by the given function $\varphi$. During computation, each parallel branch (Figure 1) outputs a stream of tuples of the form $(\hat{\varphi}(\lambda), \hat{\lambda}_i)$ (one after each sample and for each equivalence relation $\sim_l$). Each such tuple states that (for a formal statement see Theorem 1 in Section SM1.2.1), with statistical confidence $(1 - \delta)$, the probability that further sampling within $\Lambda_i$ will disprove that $\hat{\lambda}_i$ is a CAPP of $\Lambda_i$ with respect to $\sim_l$ is $< \varepsilon_l$. The algorithm in Figure 1 includes a periodic revision of the sampling policy in order to exploit the new acquired knowledge (of course at the price of resetting all counters $N_i \in [1,L]$).

3 Computing complete stratified populations for a VPH model of the HPG axis

In this section we show how we instantiated the general methodology described in Section 2 to a complex state-of-the-art VPH model of the HPG axis (called GynCycle) in order to compute a stratified set of CAPPs. We argue that our approach is based on general concepts applicable to a wide class of VPH models, e.g., those defining hormonal signalling networks.

3.1 The GynCycle model

GynCycle (Riblitte et al., 2013) is a VPH model of the human female HPG axis with a special focus on the interactions and feedback mechanisms at different stages of the menstrual cycle. The model (see Section SM2.1 for more details) defines, by means of parametric highly non-linear ODEs, the dynamics of 33 biological species (mostly hormones) having a role in the menstrual cycle (e.g., GnRH, FSH, LH, E2, P4 among the others) and the PK/PD of two pharmaceutical compounds. In particular, model inputs encode administrations of GnRH analogues that alter the menstrual cycle.

We formalised our GynCycle model as a dynamical system $S$ (Section 2.1) as follows.

Time span. Due to the model complexity, GynCycle simulations evolve not to be computed by numerical simulation. This results in both input and observation functions being bounded-time sequences of samples evenly spaced in time. To obtain robust results, we computed physiological meaningfulness metrics (Section 3.2) and phenotypes (Section 3.3) across 120 days (i.e., roughly 4 menstrual cycles), after ignoring the first 3 cycles (to get rid of any transient model behaviours, with this value being established by preliminary experiments). The time quantumbetween samples was set to 14.4 minutes (i.e., 100 samples per day) to account for the physiological time scales of the modelled signalling pathways. Hence, input and observation functions are encoded as sequences of $h = 12000$ samples, every 14.4 minutes.
Parameter space. The model counts 76 real-valued patient-specific parameters (e.g., hormone decay rates, reaction rates, stimulatory and inhibitory effects) with known bounds (Roblitz et al., 2013). By preliminary experiments we assessed that a change of parameter values of <10% yields very small changes in the resulting model trajectories (model locality). Hence, by discretising the interval for each parameter into 10 values, we produced a finitised parameter space \( \Lambda \) of size \( 10^{76} \). Although this size is still too large to be explored exhaustively. However, thanks to our informed sampling policy (Section 3.4), we were able to compute large APPs proved complete with a high statistical confidence (95%) and a small error margin (as low as \( 5 \times 10^{-3} \)).

Model inputs. Model inputs define doses for each of the two supported pharmaceutical compounds. Thus, an input time function defines a time sequence of doses administered for each of the two compounds.

Model outputs. Model outputs are non-negative real values for the \( \upsilon \in \mathbb{N}_+ \) model observables. In Section 3.5 we experiment with \( n = 4 \) observables, namely: LH, FSH, E2, P4, which are the hormones typically measured in a clinical setting, and for which we have retrospective data (Section 3.5.3).

3.2 Physiological meaningfulness

In (Roblitz et al., 2013), GynCycle has been fitted against a database (courtesy of Pfizer) comprising 20–25 measures for 4 observed hormones (E2, P4, FSH and LH) on 12 healthy women, totalling more than 1000 measurements. This activity produced a parameter assignment \( \Lambda^{(0)} \) which entails model behaviours averaging those of such 12 patients (see Section SM2.2).

In hormonal signalling pathways like those in GynCycle, all healthy humans show the same qualitative time course of such hormones. Hence, \( \lambda^{(0)} \) defines a VP that we can (and do) regard as a reference VP. Thus, we defined function \( \varphi \) (which encodes the physiological meaningfulness criteria that must be satisfied by a parameter assignment \( \lambda \) for it to be considered a VP, see Section 2.1) asking for (loose) qualitative similarity between the model observation functions under \( \lambda \) and those under \( \lambda^{(0)} \). Namely, we proceed as outlined in the following sections.

Representative portfolio of input functions. In order to derive VPs whose behaviour is meaningful also when drugs are administered, we defined a representative portfolio \( \mathcal{U} \) of 5 different input functions. Beyond the no-drug input (under which the GynCycle observation function must represent a healthy natural menstrual cycle), we considered two standard treatment strategies, consisting of daily administrations of two different doses for each of the two pharmaceutical compounds supported by the model (see Section SM2.2.1).

Physiological meaningfulness as qualitative similarity. Our function \( \varphi \) returns true on \( \lambda \in \Lambda \) (thus declaring \( \lambda \) to be a VP), if and only if the model observation functions under \( \lambda \), when subject to each of the input functions in \( \mathcal{U} \), have values always within certain physiological bounds, and can be (jointly) time-scaled and/or time-shifted (up to a certain limit) so to satisfy certain qualitative similarity metrics, when compared to the observation functions entailed by the reference VP \( \lambda^{(0)} \) under the same input. Time-shifting and scaling allow us to deal with time-alignment issues and different menstrual cycle durations, respectively.

The qualitative similarity metrics we exploited are standard (discrete-time) signal processing metrics (e.g., Vaseghi, 2000): the Normalised Zero-Lag Cross-Correlation (NZC) and the Normalised Energy Difference (NED), which we require to be, respectively, above and below certain thresholds. In our experiments, we set such thresholds to 70% and 80%, respectively. We also set limits for time-scaling and time-shifting to ±10% and 35 days, respectively. Such values (defined after preliminary experiments) are generous enough to allow us to accept model behaviours quite different from those entailed by the reference VP, but still appearing physiologically meaningful to a visual inspection.

The intuition behind and the formal definitions of our metrics, as well as technical details on how \( \varphi(\lambda) \) is actually computed (for any given \( \lambda \in \Lambda \)) are reported in Section SM2.2. Here, we just point out that such computations are quite heavy. In particular, GynCycle must be numerically simulated under each candidate parameter \( \lambda \) and each input function \( \upsilon \in \mathcal{U} \), in order to retrieve the observation function \( y(\upsilon, \lambda) \). Also, time-scaling and time-shifting issues must be evaluated before comparing our similarity metrics between \( y(\upsilon, \lambda) \) and \( y(\upsilon, \lambda^{(0)}) \). To cope with such issues efficiently, our approach envisions the solving of a constraint satisfaction problem to enumerate all possible peak alignments between the two observation functions, and the use of algorithms to compute NZC and NED between the (time-scaled and time-shifted) \( y(\upsilon, \lambda) \) and \( y(\upsilon, \lambda^{(0)}) \), for each \( \upsilon \in \mathcal{U} \).

3.3 Stratified phenotypes

Our definition of behavioural indistinguishability (i.e., same-phenotype equivalence relation) of different VPs follows an approach consistent to the one we used to decide physiological meaningfulness. However, in this case, similarity is quantitatively evaluated between the observation functions of each pair of VPs \( \iota, \lambda \). As a matter of fact, at the end of Section 3.2, already satisfy the qualitative similarity metrics thresholds against the reference VP \( \lambda^{(0)} \).

To compare two observation functions available in the form of discrete sequences of real-valued samples evenly spaced in time (as is our case), we compare the coefficients of their Discrete Fourier Transforms (DFTs) (see, e.g., Vaseghi, 2009). In particular, to define behavioural indistinguishability among VPs, we use an equivalence relation \( \sim_{\psi} \), parametric in \( \psi \in \mathcal{R} \) (the quantisation factor). Two VPs \( \lambda^{(1)} \) and \( \lambda^{(2)} \) belong to the same equivalence class (i.e., \( \lambda^{(1)} \sim_{\psi} \lambda^{(2)} \)) if and only if the DFT coefficients of their associated VPH model observation functions (for all observables and for all input functions \( \upsilon \in \mathcal{U} \)) belong to the same quantum (for a formal statement see Definition 4 in Section SM2.3). The size of quanta for DFT coefficients is inversely proportional to both \( \psi \) and the energy of the observation function of each model observable \( \upsilon \in \mathcal{U} \) under the distinguished parameter assignment \( \lambda^{(0)} \) (\( |y(\upsilon, \lambda^{(0)})|^2 \)) and its associated quantisation factor \( \psi \), which acts as a normalising factor. This is important, because the different model observables may assume values in very different ranges. In our experiments (Section 3.5), \( \lambda^{(0)} \) is the GynCycle reference VP.

Our definition of \( \sim_{\psi} \) implies (see Remark 1 in Section SM2.3) that \( \psi \) is an upper bound to the NED shown by the observation functions of any two VPs \( \lambda^{(1)} \) and \( \lambda^{(2)} \) such that \( \lambda^{(1)} \sim_{\psi} \lambda^{(2)} \), for any model observable \( \upsilon \in \mathcal{U} \) and input function \( \upsilon \in \mathcal{U} \). Thus, by considering \( L \) increasing values for \( \psi \), \( \psi_1 < \cdots < \psi_L \), \( L \in \mathbb{N}_+ \), we define \( L \) equivalence relations \( \sim_{\psi_1}, \cdots, \sim_{\psi_L} \) such that VPs in larger behavioural indistinguishability classes as their associated quantisation factor increases (stratified phenotypes). In our experiments (Section 3.5), we choose \( L = 7 \) and an increasing set of \( 7 \) values for \( \psi \) (see Table 1), where \( \psi_1 \) is such to place all generated VPs into a single equivalence class.

Indeed, value \( \psi \) turns out to be a very loose upper bound for the NED between VPs belonging to the same equivalence class. This is because it does not take into account the fact that all our VPs are known to satisfy the physiological meaningfulness criteria of Section 3.2 (qualitative similarity with respect to the behaviour of the VPH model under parameter \( \lambda^{(0)} \)). In particular, since such criteria depend on optimal time-shifts and time-stretches sought for each single VP, our bound to the NED cannot exploit such knowledge and needs to stick to the worst-case. To this end, in our experimental analysis, we also compute, by means of auxiliary hypothesis testing–based SMC tasks (along the lines of our main algorithm of Section 2.2.1, with error margin 1% and confidence ratio 5%), the actual
maximum NED between VPs belonging to the same equivalence class of each stratum (see Table 1).

3.4 Sampling policy and parallel computation

Like many VPH models, GynCycle is organised in several components, one for each of the modelled hormones. Changing the values of the elements of the parameter vector occurring in a few components typically changes the overall model dynamics only partially.

This key observation is at the heart of our sampling policy. In order to draw, with high probability, a parameter assignment that proves to be a VP, we exploit the knowledge acquired in the past iterations, in terms of the parameter assignments that already proved to be VPs. Namely, let \( \hat{\Lambda} \) be the set of VPs already discovered (population of known VPs). Our sampling policy draws a random parameter \( \lambda \) by changing uniformly at random the elements occurring in \( \mu \in \hat{\Lambda} \) model components (chosen uniformly at random) from a parameter \( \hat{\lambda} \) chosen uniformly at random from \( \hat{\Lambda} \) (if \( \hat{\Lambda} \) is empty, then \( \lambda = \lambda^{(0)} \)). Value of \( \mu \) is drawn from a Zipf’s distribution (i.e. \( \mu \sim \alpha^{-\gamma} \), where \( \alpha \) is a normalisation factor), in order to draw with high probability small values. In our experiments we set \( \gamma = 3 \) so that the expected value for \( \mu \) is about 1.11.

The sampling policy is periodically revised by updating \( \hat{\Lambda} \) with the new discovered VPs. However, in order to avoid too frequent policy updates (which would resort in an immediate reset of the consecutive failure counters, see Section 2.2.1), set \( \hat{\Lambda} \) is updated only every a given number \( N \) of samples. In our experiments we chose \( N \) such that experiencing \( N \) consecutive failures to find a new VP (regardless of its phenotype) would allow us to conclude, with statistical confidence \((1 - \delta) = 99\%\), that the probability that additional VPs will be found by further sampling is less than \( \varepsilon = 1 - 0.99^{1/N} = 5 \times 10^{-5} = 0.0005\% \). This results in \( N = 59914 \).

For the above sampling policy to work on top of a slicing of the parameter space \( \Lambda \) to be processed in parallel, it is enough to ensure that \( \lambda^{(0)} \) belongs to all slices. This was done by defining our (initially continuous) parameter space finitisation as a grid having \( \lambda^{(0)} \) as one of its vertices, and by defining the \( k \) slices by bisecting \( \Lambda \) on values \( \lambda^{(0)}, \ldots, \lambda^{(0)} \) (for any subset of coordinates \( i_1, \ldots, i_r \) within \([1, 76]\), thus defining \( k = 2^r \) slices \( \Lambda_1, \ldots, \Lambda_k \) all containing \( \lambda^{(0)} \)). In our experiments we chose \( r = 7 \) random coordinates, hence \( k = 128 \).

3.5 Experimental results

Here we present our results on GynCycle. In Section 3.5.1 we show the APPs we computed, in Section 3.5.2 we analyse the behaviour of our sampling policy, and in Section 3.5.3 we perform a qualitative and quantitative evaluation of the representativeness of our populations with respect to retrospective clinical data (86 medical cases courtesy of Hannover Medical School, University Hospital of Lausanne, and Pfizer).

3.5.1 Computed APPs

We ran our SMC-based algorithm on a parallel HPC infrastructure (the Marconi cluster at Cineca, Italy) with the settings defined above, in order to compute the stratified APPs as defined in Section 3.3. Confidence ratio \( \delta \) was set to 0.05.

The computation was stopped after around 60 days. In total, our algorithm sampled 414 245 648 parameters (simulating GynCycle for 7 months on each of them and for each of the input functions in the representative portfolio described in Section 3.2). Overall, 48 302 264 parameters were declared to define VPs.

Table 1 lists the sizes of the 7 computed APPs. The bottom line refers to the entire population of VPs, \( \hat{\Lambda} \) (which is an APP with respect to equivalence relation \( \approx \)).

We decided to terminate our (any time) computation when we achieved \( \varepsilon = 5 \times 10^{-5} \) for all slices on the top three strata. This means that (see Section 2.2.1 and Theorem 1 in Section SM1.2.1), with statistical confidence \( 1 - \varepsilon = 95\% \), the probability that further sampling (in any single slice) would disprove that such top three APPs are indeed CAPPs is below the error margin \( 5 \times 10^{-5} \).

For the above sampling policy to work on top of a slicing of the parameter space \( \hat{\Lambda} \) to be processed in parallel, it is enough to ensure that \( \hat{\Lambda} \) is about 1.11.

As for the other strata, the table reports minimum, maximum and average error margins across the \( k = 128 \) parallel processes (one per slice) at the time of termination of our any-time computation. Since the exploration of each slice is an independent process, the \( k \) error margins for each stratum can be quite different, as the value for \( \varepsilon \) for a given slice depends on the time when the last VP belonging to that slice was generated.

Also, when we terminated our experiment, a new VP (a phenotype known to the top three strata) was just generated. Hence, the max \( \varepsilon \) for the population \( \hat{\Lambda} \) consisting of all VPs (bottom line of Table 1) is 1.

Figure 2a) shows the trajectories of the GynCycle observables under the VPs belonging to the computed APPs for all strata except the extreme two. It can be seen that, despite the number of VPs greatly reduces at higher levels of our stratification, all APPs retain full representativeness of the entire spectrum of possible behaviours.

A final note is in order. Although 60 days could appear an unusually-long time for a computation (especially if compared to the time typically needed by classical model fitting tasks), this is a one-time activity for the input VPH model, and can be sped-up almost arbitrarily by using a higher number of parallel processes (e.g., using 1280 processes—which is perfectly feasible in today’s infrastructure-as-a-service platforms—with groups of 10 processes jointly exploring each of our 128 slices, would have required just 6 days). Indeed, once a population of VPs for a given model has been computed, it can be used to carry-out multiple ISCT (i.e., for different treatment strategies or medical devices). Each ISCT can be carried-out on the most appropriate stratum of VPs, depending, e.g., on the chosen trade-off between budgeted computational effort and required behavioural granularity of the VPs recruited for the trial. Also, more sophisticated approaches can be exploited, e.g., iterative deepening within the stratification of phenotypes (guided by simulation results) searching for a VP showing a failure of the candidate treatment or medical device (a counter-example, see, e.g., Mancini et al., 2013).

3.5.2 Sampling policy behaviour

Our informed sampling policy was able, on average, to find (within our 128 slices) an admissible VP every 86 attempts (average success rate: 1.17%). This is to be compared to a uniform (non-informed) sampling policy, which was unable to discover a single VP after 50 million attempts.

Figure 2b) shows the error margin achieved by our informed sampling policy during generation of \( \hat{\Lambda} \) (\( \approx \), i.e., the APP associated to the smallest value of \( \varepsilon \) (see Table 1) for which we reached an error margin of \( 5 \times 10^{-5} \) for all slices. The plot shows the values for the error margin reached by each of the 128 parallel computations (light curves) when discovering each of its \( x \) axis, thus disproving that the current APP was indeed a CAPP.

Values for \( x \) have been normalised into percentages of the total number of the VPs discovered by each parallel computation. We note that the average error margin (dark curve) lies for most of the time at values one order of magnitude higher than the value we chose to terminate our experiments.

<table>
<thead>
<tr>
<th>id</th>
<th>( \mu )</th>
<th>APP size</th>
<th>mean</th>
<th>avg</th>
<th>max</th>
<th>min</th>
<th>NED</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>16 200</td>
<td>1</td>
<td>( 5 \times 10^{-5} )</td>
<td>( 5 \times 10^{-5} )</td>
<td>( 5 \times 10^{-5} )</td>
<td>163.23%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8100</td>
<td>104</td>
<td>( 5 \times 10^{-5} )</td>
<td>( 5 \times 10^{-5} )</td>
<td>( 5 \times 10^{-5} )</td>
<td>434.36%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4000</td>
<td>2562</td>
<td>( 5 \times 10^{-5} )</td>
<td>( 5 \times 10^{-5} )</td>
<td>( 5 \times 10^{-5} )</td>
<td>106.74%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2700</td>
<td>43 941</td>
<td>( 5 \times 10^{-5} )</td>
<td>( 6.75 \times 10^{-5} )</td>
<td>( 4.51 \times 10^{-5} )</td>
<td>84.70%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1800</td>
<td>213 219</td>
<td>( 5.09 \times 10^{-4} )</td>
<td>( 2.36 \times 10^{-2} )</td>
<td>1</td>
<td>59.09%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>2 1 367</td>
<td>10</td>
<td>( 3.25 \times 10^{-4} )</td>
<td>( 8.33 \times 10^{-3} )</td>
<td>1</td>
<td>48.02%</td>
</tr>
<tr>
<td>1</td>
<td>—</td>
<td>4 830</td>
<td>264</td>
<td>( 9.7 \times 10^{-5} )</td>
<td>( 1.81 \times 10^{-7} )</td>
<td>1</td>
<td>9.97%</td>
</tr>
</tbody>
</table>

Table 1. Stratified GynCycle APPs. Statistical confidence: 95%.
(e.g., those based on the relative entropy of two probability distributions or the similarity of their momenta) cannot be, by definition, valid in our case.

To assess the representativeness of our APP with respect to the available datasets, we then proceed at computing a deterministic measure of coverage, by assessing the percentage of health records for which there exist a VP in our APP exhibiting a good enough fit. Such measure is defined in terms of a given upper bound of a standard error metric, the Average Normalised Mean Absolute Error (aNMAE).

Full details on how we formalise each health record in our datasets and on how we compute the aNMAE of each VP with respect to it are delayed to Section SM3. Here, we comment on Figure 3(b), which shows the coverage of our three datasets as a function of the aNMAE, as resulting from our analysis. The figure shows that most health records are covered by our population within small aNMAE values. Namely, the totality of the Pfizer, Hannover, and Lausanne medical records are covered within aNMAE 15%, 20%, and 35%, respectively. As for the latter dataset, 90% of the cases are actually covered within an aNMAE of just 20%.

4 Conclusions

In this paper we presented methods and software to compute a complete and stratified population of pairwise distinguishable VPs for a given quantitative model of the human physiology (plus drugs PK/PD). The availability of such populations is a key enabler for iSTC and model-based therapy design and optimisation (see, e.g., Mancini et al., 2018; Sinisi et al., 2020). Our approach is especially designed for complex (e.g., non-linear stiff ODE-based) parametric non-identifiable VPH models that cannot be analysed symbolically or integrated in closed form, but must be numerically simulated. To this end, our algorithm runs a global search on the space of model parameter assignments, guided by statistical model checking and hypothesis testing, and exploiting suitable biological and medical knowledge elicited from experts to recognise physiologically meaningful behaviours and different phenotypes, as well as structural knowledge of the model to intelligently drive the search via an informed sampling policy. Our algorithm can be stopped at any time, since it continuously provides an upper bound (correct with a user-defined confidence level) to the probability that further computation will discover new phenotypes.

We proved the effectiveness of our algorithm on a state-of-the-art non-identifiable ODE-based VPH model of the female HPG axis, by generating a population of 4 830 264 VPs stratified into 7 levels (at different granularity of behaviours), and assessed its representativeness against 86 retrospective health records.

References

Fig. 3. (a) Qualitative and (b) quantitative validation of our GynCycle population against clinical data.


