Received: 12 July 2022 | Accepted: 9 August 2022

DOI: 10.1002/psp4.12858

ARTICLE

Model enhanced reinforcement learning to enable precision dosing: A theoretical case study with dosing of propofol

Benjamin Ribba | **Dominic Stefan Bräm** | **Paul Gabriel Baverel** | **Richard Wilson Peck**

Roche Pharma Research and Early Development (pRED), F. Hoffmann La Roche Ltd., Basel, Switzerland

Correspondence
Roche Pharma Research and Early Development (pRED), F. Hoffmann La Roche Ltd. Grenzacherstrasse 124, 4070 Basel, Switzerland.
Email: benjamin.ribba@roche.com

Present address
Dominic Stefan Bräm, University Children's Hospital Basel, Spitalstrasse 33, 4056 Basel, Switzerland.

Paul Gabriel Baverel, Molecular Partners AG, Wagistrasse 14, 8952, Schlieren, Switzerland.

Richard Wilson Peck, Department of Pharmacology & Therapeutics, University of Liverpool, UK.

Abstract

Extending the potential of precision dosing requires evaluating methodologies offering more flexibility and higher degree of personalization. Reinforcement learning (RL) holds promise in its ability to integrate multidimensional data in an adaptive process built toward efficient decision making centered on sustainable value creation. For general anesthesia in intensive care units, RL is applied and automatically adjusts dosing through monitoring of patient's consciousness. We further explore the problem of optimal control of anesthesia with propofol by combining RL with state-of-the-art tools used to inform dosing in drug development. In particular, we used pharmacokinetic-pharmacodynamic (PK-PD) modeling as a simulation engine to generate experience from dosing scenarios, which cannot be tested experimentally. Through simulations, we show that, when learning from retrospective trial data, more than 100 patients are needed to reach an accuracy within the range of what is achieved with a standard dosing solution. However, embedding a model of drug effect within the RL algorithm improves accuracy by reducing errors to target by 90% through learning to take dosing actions maximizing long-term benefit. Data residual variability impacts accuracy while the algorithm efficiently coped with up to 50% interindividual variability in the PK and 25% in the PD model's parameters. We illustrate how extending the state definition of the RL agent with meaningful variables is key to achieve high accuracy of optimal dosing policy. These results suggest that RL constitutes an attractive approach for precision dosing when rich data are available or when complemented with synthetic data from model-based tools used in model-informed drug development.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Precision dosing is uncommon in healthcare, even for narrow therapeutic index drugs, partly because of the complexity of developing informative dosing algorithms to be tested during drug development. Reinforcement learning (RL) combined with pharmacokinetic-pharmacodynamic (PK-PD) modeling shows promise for the development of clinically meaningful precision dosing algorithms.

WHAT QUESTION DID THIS STUDY ADDRESS?

The study shows an application of RL when coupled with PK-PD modeling for real-time control of general anesthesia with propofol and investigate the robustness of model enhanced RL to deliver improved precision dosing algorithms. Common drug development challenges, such as small patient populations in early clinical development, partial definition of patient's drivers of response, and presence of different levels of variability in the data were explored.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

The additional benefit of coupling RL with PK-PD modeling to improve dosing precision for propofol was demonstrated compared to conventional dosing strategies. Error to target can be reduced by 90% in a theoretical scenario of absence of measurement errors and low interindividual variability. Definition of the learner state is key and residual error can have significant impact on the efficiency of the method due to potential state misspecification.

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

The nexus between artificial intelligence techniques and PK-PD modeling is further expanded with the introduction of model enhanced RL as a concept with promising properties. This method can constitute a basis for future integration of high frequency and multidimensional patient-level data to identify the right dose for every patient.

INTRODUCTION

Precision dosing is a key enabler for precision medicine by taking each patient and their disease's individual characteristics into account to find the right dose for them rather than assuming everyone is the same and needs the same dose.[1] Precision dosing has recently gained momentum in the clinical pharmacology literature[2] but its application in clinical practice or in the broader drug development context remains rare. The "one-size-fits-all" paradigm and the incentive to formulate uniform label claims in terms of dosing regimen[3] are confining the applicability of precision dosing to very few approved drugs. Yet, precision medicine is gaining momentum and key stakeholders are beginning to recognize its importance.[4] A recent US Food and Drug Administration (FDA) analysis suggested that quite a high fraction of approved drugs were amenable to response-guided dosing[5] and precision dosing is likely to help developing and monitoring drugs with narrow therapeutic windows, such as many immunomodulatory agents.

Precision dosing requires one or both of an understanding of how pretreatment or disease characteristics impact the choice of dose and a target level of effect to guide dose adjustment. The target may be a clinical response, biomarker, or a drug concentration, and today, dose adjustments are usually intuitive or sometimes guided by simple algorithms coded in dosing tables such as those in the labels for drugs such as omalizumab[6] or intravenous immunoglobulin replacement therapy[7] or the use of pharmacokinetic (PK) models as defined in vancomycin dose adjustments.[8] For such model-based precision dosing, pharmacometricians use Bayesian inference techniques and therapeutic drug monitoring to individualize treatment based on an assumed exposure-response model. Based on a prespecified PK and pharmacodynamic (PK-PD) model, patient-level data are used to estimate parameter values with the highest likelihood to represent patient's current (observed) and future (unobserved) PK-PD trajectory and individual model predictions are used to investigate optimal dosing regimen.

Pharmacokinetic-pharmacodynamic model-enhanced reinforcement learning (RL) can be seen as a generalization of these approaches. First, it allows for the incorporation of multidimensional PK-PD drivers into dosing algorithms. Second, it allows the precision dosing policy to be defined dynamically, allowing RL to continuously learn, by itself, the "dosing table" through real and simulated experience. Figure 1 shows an illustration of how we foresee RL in the precision of tomorrow (right panel) compared to how precision dosing is implemented today (left panel).

Reinforcement learning has shown preliminary promise when applied to the problem of defining dose and precision dosing algorithms. In the field of oncology, for instance, interesting analyses developed the concept showing powerful application in some specific therapeutic areas, as for reduction of tumor burden following chemoradiotherapy[9-12] or managing neutropenia in patients with cancer treated with chemotherapy.[13] We have herein studied precision dosing of propofol to explore methodological considerations of the coupling of RL with population modeling to enable optimal dosing regimens. Propofol is used for both induction and maintenance of general anesthesia. Especially when used for maintenance of anesthesia as part of total intravenous anesthesia regimens, it is recommended that patients be monitored using processed electroencephalogram (EEG)[14] with propofol dose adjusted to maintain the bispectral index (BIS), derived from the EEG signal, in the range of 40–60. Monitoring can be manual or with closed loop systems where the anesthesiologist determines the target BIS level and dose adjustments are made automatically in response to deviations of measured BIS from the desired target. For research purposes, different (model-informed) precision-dosing tools have been already tested on patients and some are actually routinely used in the clinic. For example, model-based closed-loop of propofol administration using BIS as control variable,[15] online Bayesian forecasting of PDs to improve anesthetic drug titration,[16,17] closed-loop control of administration of vasoactive drugs to control blood pressure,[18] and target controlled infusion for propofol administration.[19,20] Some meta-analyses of comparative

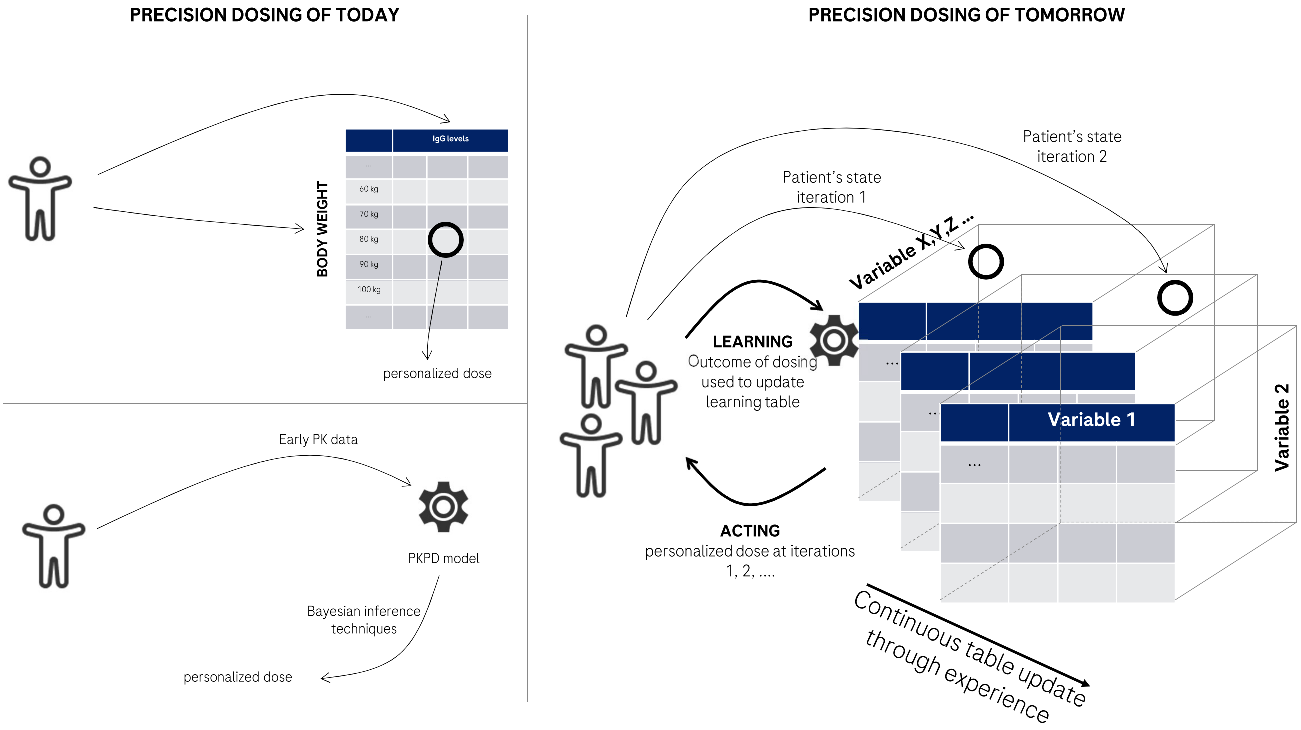
PRECISION DOSING OF TODAY
IgG levels
BODY WEIGHT
...
60 kg
70 kg
80 kg
90 kg
100 kg
...
personalized dose
Early PK data
PKPD model
Bayesian inference techniques
personalized dose
PRECISION DOSING OF TOMORROW
Patient's state iteration 2
Patient's state iteration 1
Variable X,Y,Z ...
LEARNING
Outcome of dosing used to update learning table
...
...
Variable 1
...
Variable 2
ACTING
personalized dose at iterations 1, 2, ...
Continuous table update through experience

FIGURE 1 Aspirational comparison of attributes and processes informing precision dosing presently and in the future. Today's application of precision dosing uses two-dimensional predefined table to select doses based on the patient's attributes or PK-PD models and therapeutic drug monitoring through Bayesian inference (left panel). PK-PD model-enhanced reinforcement learning is seen as a generalization of the precision dosing problem where the table with optimal dosing is constantly updated with outcomes of real and simulated applied dosing and can accept high dimensional entries as patient's attributes (right panel) creating a feedback loop process. PD, pharmacodynamic; PK, pharmacokinetic.

studies suggest that closed loop systems, such as RL, deliver better control of the BIS than does manual control, which may be associated with fewer adverse effects and improved clinical outcomes.[21] Other analyses suggest little difference in clinical effects[22] although they agree that use of closed loop systems reduces variability in the achieved BIS and reduces the total dose of propofol required to maintain anesthesia.

Herein, our focus is on the methodology and not on the application. We indeed neither claim we generate new advancements in the optimal dosing of propofol, nor that these findings could be translated to real-life applications. However, we believe methodological progresses are needed to broaden feasibility of precision dosing beyond such a specific example. For instance, there is a need for method development to take advantage of highly dimensional data from wearable devices, or digital health technology tools in general. Such tools have the potential to generate patient health-related data with much higher granularity than what is possible with classical clinical research tools.[23] Ideally, progress in methodologies will enable the integration of such data to design efficient precision dosing techniques. Herein, we selected propofol dosing as a study case mainly to illustrate and discuss the potential of model-enhanced RL through a well-characterized drug's PK-PD properties. Indeed, a specific aspect of interest for the clinical pharmacology community is the potential added value of coupling such approaches with PK-PD modeling when confronted to real-life challenges inherent to precision dosing.

We here assess the potential of the PK-PD model-enhanced RL is through the use of synthetic data. We used simulations to investigate the suitability of RL to implement precision dosing in clinical settings and drug development. More specifically, we investigated the performance of RL on three attributes typically encountered when stating a dose regimen problem.

The first challenge in clinical practice and drug development is that not all possible doses or dosing scenarios can be experimented in patients. This limitation constrains the algorithm, as experiences that did not happen cannot generate any learnings. To compensate for this problem, we designed a technique exploring state-action pair space without actually experiencing it in real life. A second challenge is the incomplete knowledge on the underlying relationship between the dose and PD or disease biomarkers. In early development of a new drug, most biomarkers are in a research or translational phase and lack clinical validation. This naturally leads to limits in defining the exact state of the agent. The third challenge is variability in the data, in particular interindividual variability and residual errors that are inherent to the collection and analysis of pharmacology data.

METHODS

General introduction to reinforcement learning

Reinforcement learning designates a set of decision-making algorithms mimicking learning by interactions with environment, similarly to how it occurs in animals and human. The goal of the method is to identify which best actions to take in given situations (i.e., map actions to situations) to maximize long-term return.[24] RL can be formalized with Markov Decision Process (MDP), a framework particularly attractive because of its applicability to a wide range of optimization problems. MDP integrates key elements of a learning agent interacting with its environment over time to achieve a goal. The agent who can take actions (e.g., to dose or not in our context) and interact with an environment which can feedback a reward based on the action taken and a state which characterizes the agent and constitutes the basis for the decision to be taken. Agent, environment, state, and reward constitute the core attributes of an MDP and define the formalism of RL. The Markovian property holds if the current state is sufficient to determine the optimal successive actions maximizing the long-term rewards. In other words, the sequence of optimal future actions does not depend on what happened before the current state when a decision needs to be taken.

The aim of the algorithm is to identify an optimal policy (i.e., optimal succession of state-action pairs that generate the highest cumulated reward [sum of rewards received following each action taken]). It is not enough to try to get the best reward at each individual step of a process. In fact, most interesting problems arise when an action has not only an immediate consequence in terms of reward but also does affect the next situation and, consequently, the next rewards. What actually matters is the long-term cumulated reward. In an optimal policy, some of the actions might not be the ones leading to the highest instantaneous reward but the ones maximizing rewards in subsequent actions. As an analogy, a tennis player can deliberately choose to lose a game on the opponent's service to save energy and focus on the next game where he/she will serve for the set. An RL problem can be solved by means of value (or action-value) functions, usually called “Q” holding the expected cumulated reward from a given state (or state-action pair). If Q is known, then the solution of the problem is easy and consists of identifying which actions maximize Q in any situation or state. However, while defining the reward is often easy, estimating return (so Q) is much more difficult as it must be estimated from the sequences of observation that a learning agent makes, in other words, it has to be estimated through experience.

Delayed reward and experience are the two most distinguishing features of RL.[24] Algorithms exist to estimate the Q function. Here, we have used Q-learning temporal differences (see Appendix S1 for more details).

For a realistic environment, such as robotic platforms or dosing of a therapeutic intervention, RL algorithms must be constrained to avoid a dosing policy unsafe for the environment.[25] In this study, we defined a safe environment by constraining RL algorithm to always apply a dose when BIS is above 60 and interrupt dosing when BIS is below 50 while our target is a BIS of 55 (see below for further details). By constraining the actionable window of RL, we reduce the size of the state space. Of course, the minimal and maximal values of the window can be changed. In our view, the primary objective of this constraint is the prevention of harmful therapeutic intervention or unnecessary additional doses but should not require excessive degree of expert knowledge which could lead to a customized solution.

The algorithm is then used to remain as close as possible to the target within the prespecified safety window.

In a healthcare setting, RL aims at realizing the optimal dosing for every patient, accounting for its unique disease trajectory, its baseline credentials, and the dosing actions of a practitioner in light of the therapeutic window of the drugs administered. Recently, we published a minireview on the concept of RL applied to clinical pharmacology.[26] We believe that RL is a technology worth exploring to support precision dosing given its ability to integrate multidimensional biological and clinical data and (not explored herein) relax the need for an assumed pharmacology model when sufficient data becomes available. More details on the implementation of the RL algorithm are presented in Appendix S1.

Differences with common pharmacometric and other supervised approaches

With respect to methods commonly known in pharmacometrics—namely therapeutic drug monitoring or Bayesian inference of PK-PD model parameter individualization—RL represents a very different paradigm. Although, with Bayesian forecasting, individual data can be used (even on real time) to adapt a dosing regimen, there is no feedback from the experience to the (structural) model from which the dose individualization is derived, and consequently the optimization process, is not adaptable. In RL, and in theory still (i.e., beyond the remit of the present study where we rely on a pre-existing PK-PD model), experience is generated out of the decision rule being learned; and this experience feeds back on the learning process. Translated into pharmacometric notions, this is similar to learning simultaneously the PK-PD model from the data and the optimal dosing regimen based on this model, whereas the data are being collected. RL brings model building and optimization in one unique step. In addition to that, the learned policy is valid for all individuals and provides individualized dosing regimen based on each patient trajectory throughout the different states. RL constitutes a very different paradigm to other machine learning approaches, in general, as it learns based on online feedback from its environment and avoids direct instruction on what to achieve, thus providing an interesting avenue for problems arising from iterative clinical decisions.

Pharmacological model and simulations

The pharmacological model used as a simulation engine to generate data in the present work recapitulates the propofol mechanism of action and effect on sedation in the intensive care unit as implemented in ref. [27] from an original version presented in ref. [28]. The model is a four-compartment model described by a set of linear ordinary differential equations accounting for the distribution of the drug in various compartments, complemented by an algebraic equation linking the PDs to the concentration of propofol at the site of action modeled as an effect site compartment. The PD end point predicted is the bispectral index or BIS, which is one of the metrics to monitor the patient's level of consciousness. The PK-PD model is presented in Figure 2a left (adapted from ref. [27]). The PK model is a four-compartment linear model and the PD equation (BIS) is governed by an algebraic equation as a function of propofol concentration in the effect compartment. Equations, parameters' definitions, and values are provided in the Appendix S1. Note that the BIS could be governed by a more realistic nonlinear equation as in ref. [29].

Before the drug administration, the patient has a BIS close to 100 (full consciousness). The optimization problem can be formulated as finding the optimal dosing regimen, for each patient, which allows the BIS to be the closest to a target value ($\text{BIS}_{\text{target}} = 55$). Figure 2a right shows a typical simulation of drug concentration and BIS time course with a dosing regimen called "standard." This dosing regimen consists of giving a unit of dose at each time step where the BIS is above the $\text{BIS}_{\text{target}}$ while refraining from dosing during the time

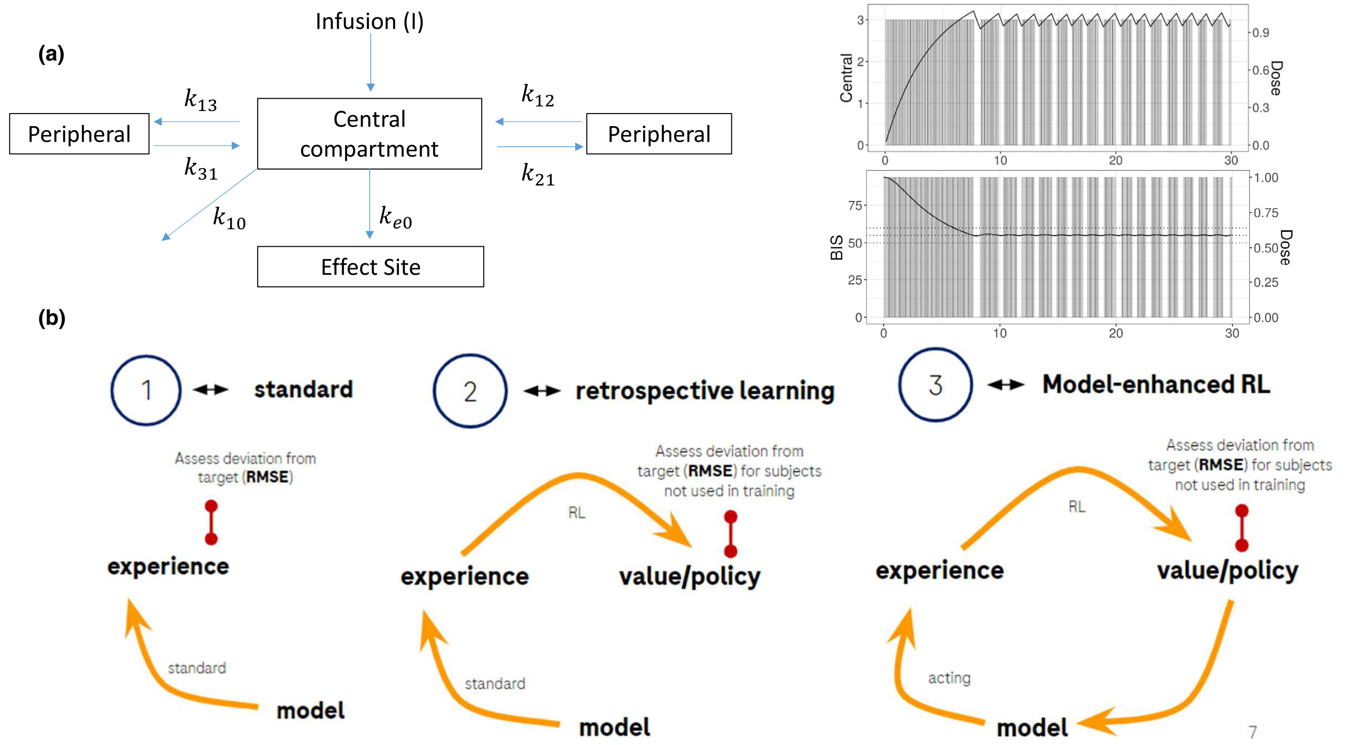

FIGURE 2 (a) Schematic representation of the PK-PD model used for data generation and prediction in some of the scenarios explored (left). The structural models as well as parameter values were taken from ref. [27] and reported in Table S1. PK-PD simulation for a typical individual receiving the standard dosing regimen which consists of applying the dose when BIS is above the target. In the figures, dose applications are shown with vertical gray lines (right). (b) Overview of the three approaches applied in the manuscript. 1 – Standard, consisting of using the model to simulate a virtual population on which the standard dosing regimen is applied (left). 2 – Reinforcement learning (RL), consisting of learning the dosing policy through RL based on the retrospective analysis of data (experience) simulated by the model. 3 – Model-enhanced RL, consisting of learning the policy through experience. For approaches 2 and 3, the estimated optimal dosing policy was applied on a new virtual population consisting of 100 patients (not used for training) and accuracy assessed through root mean squared error (RMSE). BIS, bispectral index; PD, pharmacodynamic; PK, pharmacokinetic.

steps where the BIS is below the $\text{BIS}_{\text{target}}$. The protocol results in a multiphasic BIS time profile: a first sharp decrease of BIS from its baseline value toward the target and a subsequent oscillating phase around the $\text{BIS}_{\text{target}}$, with amplitude depending on the PK-PD parameters of the model. Although the "standard" dosing regimen is an oversimplification of the current state-of-the-art for propofol precision dosing, we used it throughout this study to benchmark the performance of the RL approaches. Conscious that this oversimplification might introduce bias in the interpretability of the results, the findings should not be seen as potential application of propofol precision dosing improvement but as methodological considerations.

We used the model in three methods:

1. Standard: The PK-PD model is used to generate a benchmark scenario where the virtual patients were applied the standard dosing regimen (see Figure 2b, method 1).
2. "Retrospective learning or "model-off": The PK-PD model is used to simulate a virtual population of patients treated with the standard dosing regimen. Only this population is used to train the RL algorithm in finding optimal dosing regimen (see Figure 2b, method 2).
3. Model-enhanced RL or "model on": The RL algorithm embeds in its core the PK-PD model to predict the result of any given dosing action. In doing so, the algorithm can explore any state-action pair and the optimal dosing policy is learned through (virtual) experience (see Figure 2b, method 3).

We evaluated the accuracy of the different methods through a measure of distance to the target (i.e., root mean squared error [RMSE] see Appendix S1 for more details). For each evaluation, we used the same approach consisting of simulating new patients (not used for training) and applying the identified optimal policy. In addition to RMSE, we also recorded the number of

TABLE 1 Positioning of the present research throughout a description of the three methodological considerations we explore in this paper, namely incomplete dose ranging, incomplete knowledge and observations of disease trajectory, and variability/error

	Exploration of state-action pair	State definition	Variability and error
Train a computer to play a video game	In a learning phase, the computer can experiment any action: intense exploration of the space of state-action pairs	The computer has a complete and univocal definition of the state of the agent. Nothing outside the elements of the game (e.g., board, pieces, ...) can be considered part of the state definition	The result of an action is usually known without uncertainty
Train a computer to individualize a dosing regimen	*Challenge 1* Some of the actions (dosing) cannot be taken (due to benefit–risk balance; ethical considerations; or finite patient population)	*Challenge 2* The “state” in the dosing problem is by definition not completely known. What are the patient's variables to record for informing what best dosing action to take? Answering this question requires a holistic understanding of disease and therapeutic action	*Challenge 3* The result of a dosing action is feedback with inherent imprecision and error (i.e., uncertainty)
Avenues for RL implementation to solve dose optimization problem	*Idea 1* Embedding a PK-PD model simulation engine into the RL algorithm to explore any state-action pair (referred as the “model-enhanced RL” method in the paper)	*Idea 2* Extending the state definition (e.g., we illustrated the combining of 2 temporal adjacent PD observations and the combining of PD with PK variables)	*Idea 3* Uncertainty can be compensated through the extension of the state definition and through increasing the number of virtual patients used for training

Note: For each of the methodological considerations, we draw an analogy to a video game where RL has been shown to achieve good control for challenging problems. We also describe the avenues for addressing the methodological challenges that we illustrated through simulation in the core of the paper.

Abbreviations: PD, pharmacodynamic; PK, pharmacokinetic; RL, reinforcement learning.

doses applied in the optimal policies as well as the cumulated reward.

The virtual population in all methods has, for the most part of the present study, an arbitrary size of 100 individuals for which we sampled individual parameters from a lognormal distribution assuming an arbitrary (interindividual) variability of 10% coefficient of variation for PK parameters. For some experiments, larger populations were studied, as described in the Results section. We further tested different levels of interindividual variability and residual error in the BIS data.

RESULTS

Table 1 illustrates the methodological considerations investigated in this paper when applying RL in the context of a pharmacological dose optimization problem of propofol to trigger general anesthesia. A parallel with computer game RL is made to bring clarity to the challenges inherent to precision dosing and the proposed solutions.

Coupling RL with PK-PD modeling to achieve better precision dosing

One problem of applying RL to a clinical scenario is the impossibility of exploring some state-action pair space by trial-and-error. Benefit/risk considerations, ethical considerations, and finite population size preclude testing all possible doses. To quantify the impact of an extended exploration on the identification of the optimal dosing regimen, we compared the accuracy of a policy derived by the three methods described above: standard, retrospective learning, and model-enhanced RL.

Results shown in Figure 3 compares the performance of the RL retrospective learning approach (Figure 3b–d) versus the standard dosing policy (Figure 3a). On 100 tested patients with the standard dosing regimen, the median RMSE were 18.33, 0.3, and 0.26 BIS for the time windows 0–10, 10–20, and 20–30 min, respectively. The same median accuracy was achieved by the RL method when the initial population was including 500 patients, although the identified policy is working suboptimally for a few patients (Figure 3b). In Figure 3, we also show that the mean RMSE decreases (Figure 3c) and the mean

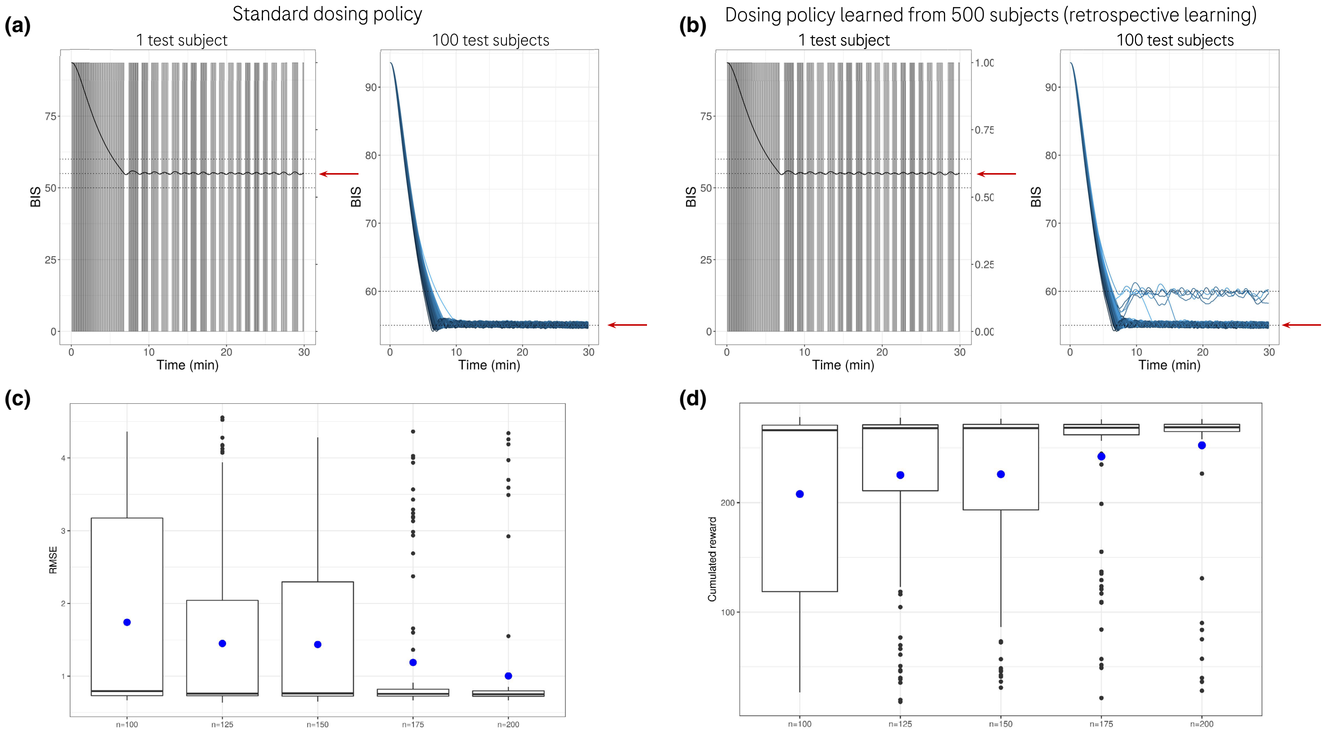

FIGURE 3 (a) BIS time course in one arbitrarily chose subject (left) and for 100 test subjects (right) with standard dosing regimen. (b) Results of reinforcement learning (RL) with initial population of 500 subjects. Root mean squared error (RMSE) (c) and cumulated reward (d) with RL based on various initial population size (100–200). Blue points denote the mean values. BIS, bispectral index.

cumulated reward increases (Figure 3d) as the size of the initial population increases. Also decreasing is the number of outliers (the subjects for which the identified policy is not optimal). Increasing the number of patients in the initial population translates in less outliers when it comes to the accuracy of the identified policy but does not modify the median RMSE or cumulated reward. In conclusion, when training with patients receiving the standard dosing regimen, the computer has no chance to learn a more effective regimen than the standard one; simply because there was not any experience with a different dosing regimen. At best, it learns the dosing regimen applied in the data used in training, and in the current case study (with 10% interindividual variability on PK parameters and no errors in the data), an initial population size of 100 patients was enough to learn the policy.

Embedding the PK-PD model in the simulation engine produces different results. In Figure 4a, we show the BIS time course on a randomly selected test subject (data not used for training) where standard dosing is applied (left) versus optimal policy with model-enhanced RL which results in removing the oscillation around the target. In terms of median RMSE, compared to the standard dosing regimen, the model-enhanced RL policy resulted in a 0.1% increase in the first 10 min (18.35 BIS vs. 18.33 for standard), but 90% decrease in the interval 10–20 min (0.03 BIS vs. 0.3), and 88% decrease in the time interval 20–30 min (0.03 vs. 0.26 for standard). Figure 4b shows the heatmap of doses used by both the standard dosing regimen (left) and the model-enhanced RL policy (right). Whereas patterns of dose application can be observed in the case of the standard regimen (representing the continuous dosing when above the target and the non-dosing when below), the dose application in the case of the RL policy is much more refined (no patterns). Importantly though, this policy leading to increased accuracy on the target did not require more doses. Indeed, the median number of doses applied in both cases (standard and model-enhanced RL) was the same: 255.

Importance of integrating multiple data dimensions for accurate precision dosing

The second methodological issue investigated is the definition of the patient's state. Our problem is defined in literature as a partially observed Markov Decision Process because, in theory, the complete state of the patient cannot be known.

Results in Figure 4a,b previously discussed were generated with state definition being composed by the current BIS and the concentration in the central compartment. In Figure 4c,d, we show the result of the model-enhanced RL when the state is defined by the current BIS (left) or two adjacent BIS, current and previous (right). Note that, in

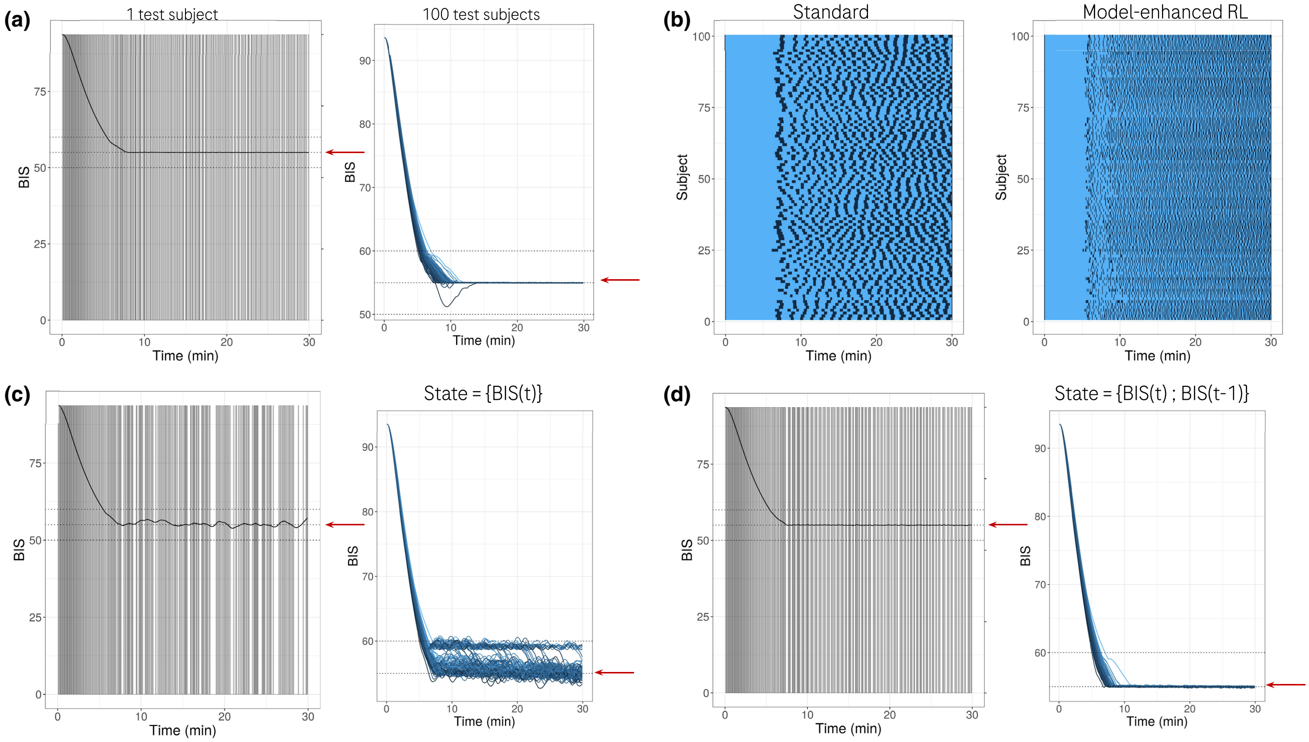

FIGURE 4 (a) PK-PD simulation in one arbitrarily chosen subject (left) and the 100 subjects from the test trial (right) when applied the model-enhanced RL. (b) Heatmap of doses applied in the standard dosing regimen (left) or model-enhanced RL (right). Light blue color indicates dosing events while dark blue no dosing. In panels a and b, the current BIS and plasma concentration define the state of the learning agent. In panels c and d, the state of the learning agent is defined differently: current BIS only (c) or two adjacent BIS (d). BIS, bispectral index; PD, pharmacodynamic; PK, pharmacokinetic; RL, reinforcement learning.

this case, the Markov property is still valid as we incorporate the previous BIS as part of the state.

When the state is defined by the current BIS, the median RMSEs were 18.36, 1.16, and 0.67 BIS for the 0–10, 10–20, and 20–30 min time windows, respectively, corresponding to an increase of 0.2%, 277%, and 162% with respect to the standard dosing regimen (18.33, 0.3, and 0.26). The median cumulated reward was 238 (min 28 and max 272), corresponding to 11% decreases with respect to the standard dosing regimen (268; min 245 and max 277).

When the current BIS and its predecessor defined the state, the median RMSEs in the three time windows were 18.36, 0.06, and 0.06, so it was unchanged for the interval 0–10 min and decreased by 80% and 77% in the intervals 10–20 and 20–30 min, respectively. The median cumulated reward was 279 (min 245 and max 288) so an increase of 4% with respect to the standard.

Effect of interindividual and residual variability on the accuracy of the estimated policy

In Figures 5 and 6, we report on performance of model-enhanced RL on virtual populations with different levels of interindividual variability and residual errors. For each scenario, the Q action-value function learning was based on a population sharing the same characteristics than the test population, although the patients from the test dataset were not included in the training. In addition, for each scenario, the results obtained were compared to the performance of the standard dosing regimen. With an additive error of 1 BIS, retrospective learning on 500 patients did worse than the standard regimen (Figure 5a,b). RMSEs increased by +0.05%, +8%, and +15% in the time windows 0–10, 10–20, and 20–30 min with respect to the standard dosing regimen. However, with the same error, the model-enhanced RL achieved RMSE changes of −0.1%, −14%, and – 7% in the same time windows with respect to the standard regimen (Figure 5c). With a combined error model (1 BIS constant and 10% proportional), the RMSEs were 0.2%, −3%, and −1% with respect to the standard dosing regimen (Figure 5d).

With 25% variability in all PK parameters, the RMSE of model-enhanced RL was +0.1%, −86%, and −85% in the 0–10, 10–20, and 20–30 min time windows (Figure 6a). For 50% variability, it was +0.05%, −60%, and – 57% (Figure 6b) whereas for 25% variability in the PD parameter (slope), these numbers were +0.1%, −69%, and −68% (Figure 6c). We also tested the method

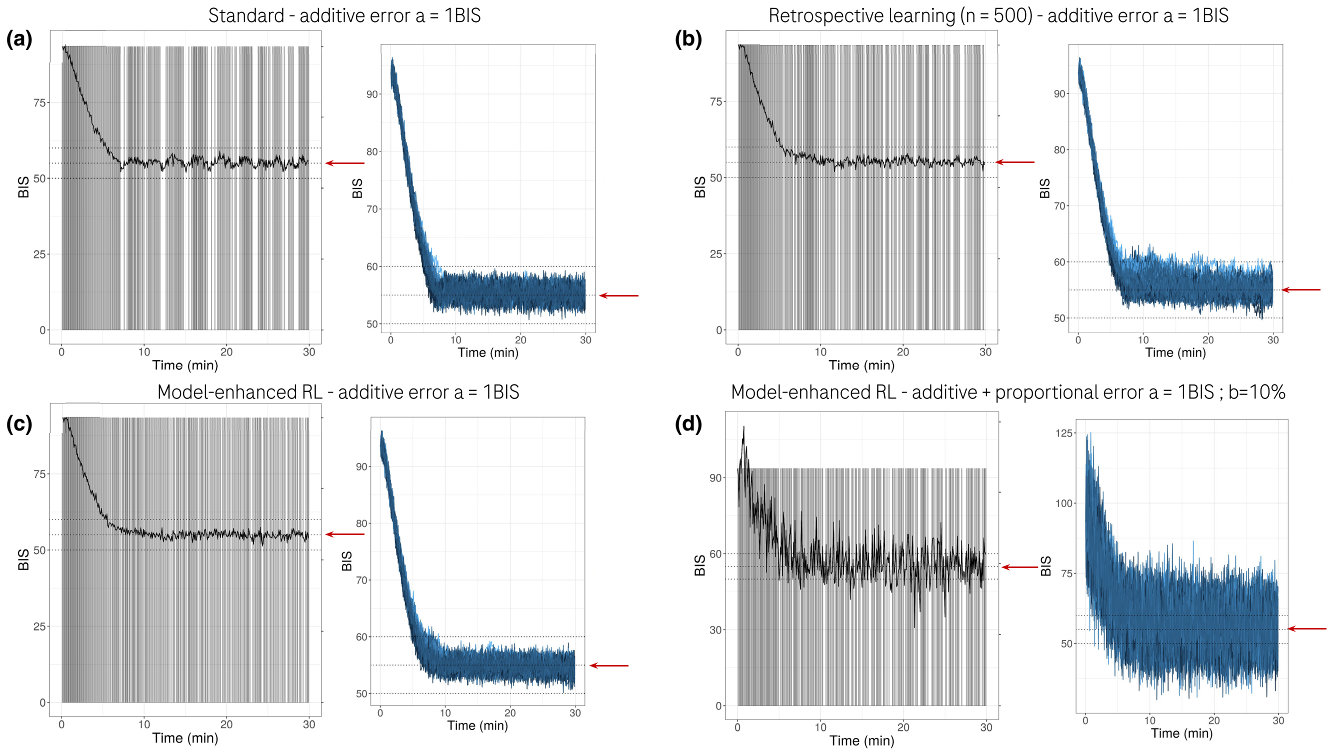

FIGURE 5 PK-PD simulation in one arbitrarily chosen subject (left) and all 100 subjects from the test population (right) in presence of residual variability. With additive error of 1 BIS, standard dosing approach (a), RL based on 500 subjects (b) and model-enhanced RL (c). (d) Model-enhanced RL with combined constant (1 BIS) and proportional model (10%). BIS, bispectral index; PD, pharmacodynamic; PK, pharmacokinetic; RL, reinforcement learning.

on a scenario combining interindividual variability in PK parameters (25%), PD (10%), and error model (constant error of 1 BIS and proportional of 10%). The model performs similarly to the standard dosing regimen with −1%, +3%, and +3% change with respect to the standard protocol for the 0–10, 10–20, and 20–30 min, respectively (Figure 6d).

Adding noise in the data can impact the algorithm decision process as it can lead to an erroneous state positioning. The agent thinks a given BIS level is realized, whereas the actual level is different. The reward being a function of the difference between the current state and the target, a spurious reward is blurring the ability of RL to establish an optimal policy with no possibility to correct for mistakes given that the noise remains constant during the learning process. When states are misperceived, it is known that RL can perform badly.[24] The reduced impact of interindividual variability versus residual variability can be explained theoretically. Interindividual variability is less of an impact because the underlying truth (the structure of the deterministic model) can still be learnt by the algorithm with low to moderate variability in the multivariate parameter space.

DISCUSSION

In this study, we report an application of model-enhanced RL for precision dosing based on a simulation example of general anesthesia following propofol administration. By coupling RL with PK-PD modeling, the results show more precise dosing with propofol over the standard dosing algorithm under varying experimental conditions mimicking common dosing problems.

This study is theoretical, as to our knowledge, propofol dosing maintenance is not a key priority for propofol precision dosing. Indeed, propofol is used for the induction of anesthesia and therefore when developing dosing recommendations, the focus is on the very early time course of BIS. Herein, we are only evaluating whether the RL approach is useful for defining dosing tables for maintenance dosing. Exploring the early induction phases could be more complicated and also more interesting for propofol, among other things, due to the hysteresis between plasma concentrations and BIS values. We have selected this example to study the behavior of model-enhanced RL in a situation where dosing optimization is non-trivial given the high dosing frequency (1 dosing each 5 s).

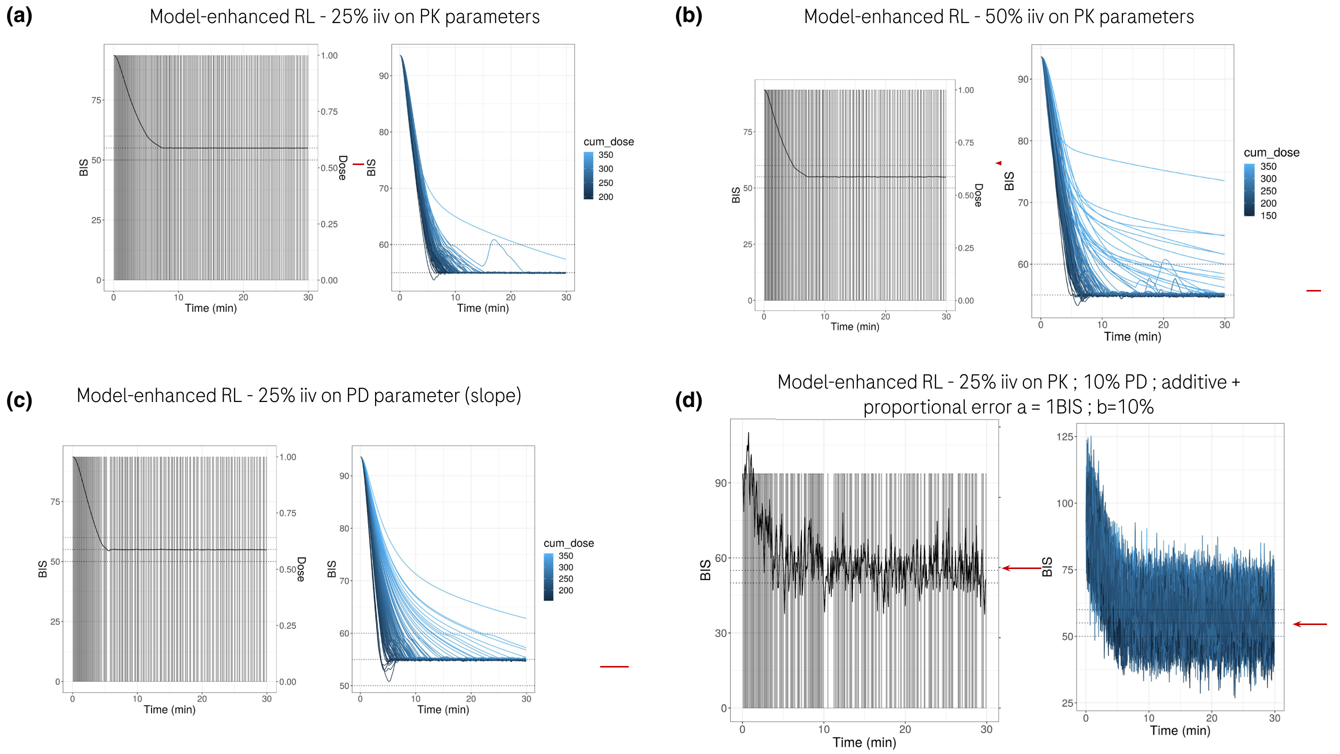

FIGURE 6 PK-PD simulation in one arbitrarily chosen subject (left) and all 100 subjects from the test population (right) in presence of residual variability. Individual variability (IIV) on all PK parameters: 25% (a), 50% (b), and 25% IIV on PD parameter (c). (d) Combined variability: 25% IIV on PK parameters, 10% on PD parameters and combined error model (constant 1 BIS and proportional 10%). BIS, bispectral index; PD, pharmacodynamic; PK, pharmacokinetic; RL, reinforcement learning.

The study presents several limitations. Regarding the potential generalization of the results to other drug and dosing scenarios, the fast onset of effect of propofol and the constant learning process throughout the course of treatment is an upside that our case example would not share with drugs typically investigated in clinical development. First, clinical data to base dosing decision are often lagging behind due to operational considerations, such as bioanalytical handling of biospecimen samples to derive PK and/or PD measurements. As such, the authors' acknowledge that propofol is a pilot study and that the current methodology proposed does not address all known challenges relevant to dosing optimization problems. In theory, however, RL can be applied to dosing problems with drugs exhibiting slower onsets and less frequent dosing regimen than propofol, as already published.[9-12]

We assumed a perfect knowledge of the underlying exposure-response following a pharmacological action. Clearly, in early clinical development, only a few scenarios would correspond to a situation where the true model is available. Our results show that increasing the patient numbers in the retrospective learning state allows RL to match standard dosing approach performance. Even with no knowledge of the PK-PD, RL can learn how to dose effectively and generate an initial precision dosing algorithm. Imperfect PK-PD models could be considered instead. For instance, a translational PK-PD model used to inform first-in-human dosing could be used as a surrogate engine if predictions are clinically relevant. In addition, the potential of reverse translation of matured clinical data to inform a disease model applied to a new compound in development targeting the same disease could be an avenue of further research. Likewise, one could consider a simple model structure that although imperfect could ensure a good exploration of the state-action pair. This structure could mature as additional clinical data become available. Our perspective includes learning the model directly from available data following artificial intelligence approaches recently published (see for instance refs. 30,31).

By investigating three key challenges relevant to dose optimization problem in clinical pharmacology (namely incomplete dose ranging, incomplete knowledge and observations of patients' disease trajectory, and error or variability in measurements), we found that RL and PK-PD modeling are complementary to address limitations of the data being analyzed. The first challenge of dose exploration led to the conclusion that model-based predictions within RL estimation is critical to compensate for the lack of exploration of the state-action pair space within a clinical trial setting. The second challenge of incomplete state definition also necessitated model-based augmented data. Extension of the state can partially compensate for lack of data. In our study, at best, retrospective learning performed similarly to the standard dosing regimen and the model-enhanced learning performs at worst similarly to the standard dosing regimen (in the presence of high residual variability).

Future research on the extension of the state to high dimensions to accommodate for high dimensional biomarker measurements is required, including the use of surrogate biomarkers of efficacy or safety of drug concentration levels. There will often be target engagement, pharmacological activity, or even efficacy markers in the periphery to inform the state of patients and that are often informative for estimating the effect of the drug. In the case of omalizumab for severe asthma, serum IgE levels at baseline are typically a target engagement marker and are used to define the recommended dosing regimen alongside baseline bodyweight in the label. Likewise, circulating proximal or distal PD markers with both pre- and post-treatment measurements could constitute useful data alongside vital signs and other standard laboratory routine measurements to inform the disease status and treatment trajectory of RL. Taken together, the improvement of RL performance when extending the state definition of patients indicates a potential of RL for aggregating and sifting through multidimensional biomarkers' data to identify optimal dosing policies.

Finally, we demonstrated that the variability in data that remains a hallmark of clinical pharmacology dose optimization problem also impacts RL performance in achieving precision dosing, in particular, the residual noise in PD measurements and/or the reward. In our specific problem, the algorithm is evaluated on its ability to reach a given PD target, directly subjected to the error of measurement. As we add errors to this target, the performance of any algorithm decays, and may never be less than the error itself.

Overall, in conclusion

- Availability of a prediction model of the consequences of drug action on the learning agents' state was key for reaching higher accuracy of the identified dosing policy with respect to the standard dosing approach.
- Defining the state in one dimension led to suboptimal results, whereas the two dimensions definition (two adjacent BIS or current BIS and plasma concentration) generated successful dosing policies.
- Interindividual variability had minimal impact, whereas random noise (residual error) had more impact due to potential state misspecification.

Further analysis could explore the potential relationships between the challenges highlighted here. For instance, could the problem of poor state definition and the problem of data variability be related? Noise in the measurements is clearly a scientific conundrum but to what extent can it be mitigated by having more measurements? We imagine a future where there are extensive digital biomarker data available to complement PK and PD standard measurements so noise in one measurement may be compensated by other less noisy data when controlled at the individual level by an automated intelligent agent, such as RL.

More research is needed to explore the potential relationship between variability in dosing regimen at the first place (in the original population) and accuracy of the optimization. In particular, the variability in the model structure or the presence of significant missing data is worth further exploration. Further investigating the impact of the RL algorithm, namely the learning and discount rate, did not appear as a priority, given that we expect their value to be linked to the specific problem and with limited potential to generate more universal learnings relevant to the precision dosing problem.

Overall, we believe this study sheds light on important methodological considerations paving the way for RL techniques as a potential tool for precision dosing. We hope its scope and content can stimulate further research in this area. Whereas RL or other machine learning (or even PK-PD modeling) methods are promising tools for precision dosing, a clear disadvantage remains the in-built complexity for actionable outputs readily interpretable. Apps, ideally directly embedded in an electronic prescribing system, are important to develop so that the physician does not have to do anything extra to find the right dose. As algorithms are developed and show promising outcomes, a parallel development of end-to-end solution enabling data collection, processing, and actionable outputs will have a positive reinforcement on developing more dosing algorithms—a virtuous spiral.

AUTHOR CONTRIBUTIONS

All authors wrote the manuscript. B.R., P.G.B., and R.W.P. designed the research. B.R. and D.S.B. performed the research. B.R., D.S.B., and P.G.B. analyzed the data.

ACKNOWLEDGMENTS

The authors wish to acknowledge Emmanuel Grenier and James Lu for methodological advice.

FUNDING INFORMATION

No funding was received for this work.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

ORCID

Dominic Stefan Bräm https://orcid.org/0000-0001-9094-8361
Paul Gabriel Baverel https://orcid.org/0000-0002-2461-5849
Richard Wilson Peck https://orcid.org/0000-0003-1018-9655

REFERENCES

1. Peck RW. Precision medicine is not just genomics: the right dose for every patient. *Annu Rev Pharmacol Toxicol*. 2018;58:105-122.
2. Bica I, Alaa AM, Lambert C, van der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther*. 2021;109(1):87-100.
3. Neely M, Onufrak N, Scheetz MH, et al. Supporting precision dosing in drug labeling. *Clin Pharmacol Ther*. 2021;109(1):37-41.
4. Maxfield K, Zineh I. Precision dosing: a clinical and public health imperative. *JAMA*. 2021;325(15):1505-1506.
5. Schuck RN, Pacanowski M, Kim S, Madabushi R, Zineh I. Use of titration as a therapeutic individualization strategy: an analysis of food and drug administration-approved drugs. *Clin Transl Sci*. 2019;12(3):236-239.
6. Administration, F.A.D. Highlights of prescribing information xolair. [cited March 21, 2022] https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/103976s5225lbl.pdf
7. Administration, F.A.D. Highlights of prescribing information gammagard. [cited February 28, 2022] https://www.fda.gov/media/70812/download
8. Rybak MJ, le J, Lodise TP, et al. Therapeutic monitoring of vancomycin for serious methicillin-resistant *Staphylococcus aureus* infections: A revised consensus guideline and review by the American Society of Health-System Pharmacists, the Infectious Diseases Society of America, the Pediatric Infectious Diseases Society, and the Society of Infectious Diseases Pharmacists. *Am J Health Syst Pharm*. 2020;77(11):835-864.
9. Yauney GS, Shah P. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. Machine Learning for Healthcare Conference, Palo Alto, CA; 2018.
10. Eastman B, Przedborski M, Kohandel M. Reinforcement learning derived chemotherapeutic schedules for robust patient-specific therapy. *Sci Rep*. 2021;11(1):17882.
11. Yazdjerdi P, Meskin N, al-Naemi M, al Moustafa AE, Kovács L. Reinforcement learning-based control of tumor growth under anti-angiogenic therapy. *Comput Methods Prog Biomed*. 2019;173:15-26.
12. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. *Stat Med*. 2009;28(26):3294-3315.
13. Maier C, Hartung N, Kloft C, Huisinga W, de Wiljes J. Reinforcement learning and Bayesian data assimilation for model-informed precision dosing in oncology. *CPT Pharmacometrics Syst Pharmacol*. 2021;10(3):241-254.
14. Nimmo AF, Absalom AR, Bagshaw O, et al. Guidelines for the safe practice of total intravenous anaesthesia (TIVA): Joint Guidelines from the Association of Anaesthetists and the Society for Intravenous Anaesthesia. *Anaesthesia*. 2019;74(2):211-224.
15. De Smet T, Struys MM, Neckebroek MM, et al. The accuracy and clinical feasibility of a new bayesian-based closed-loop control system for propofol administration using the bispectral index as a controlled variable. *Anesth Analg*. 2008;107(4):1200-1210.
16. Kuizenga MH, Vereecke HEM, Absalom AR, et al. Utility of the SmartPilot(R) View advisory screen to improve anaesthetic drug titration and postoperative outcomes in clinical practice: a two-centre prospective observational trial. *Br J Anaesth*. 2022;128(6):959-970.
17. Colin PJ, Jonckheere S, Struys M. Target-controlled continuous infusion for antibiotic dosing: proof-of-principle in an in-silico vancomycin trial in intensive care unit patients. *Clin Pharmacokinet*. 2018;57(11):1435-1447.
18. Joosten A, Alexander B, Duranteau J, et al. Feasibility of closed-loop titration of norepinephrine infusion in patients undergoing moderate- and high-risk surgery. *Br J Anaesth*. 2019;123(4):430-438.
19. Vellinga R, Hannivoort LN, Introna M, et al. Prospective clinical validation of the Eleveld propofol pharmacokinetic-pharmacodynamic model in general anaesthesia. *Br J Anaesth*. 2021;126(2):386-394.
20. Struys MM, De Smet T, Glen JI, Vereecke HE, Absalom AR, Schnider TW. The history of target-controlled infusion. *Anesth Analg*. 2016;122(1):56-69.
21. Wang D, Song Z, Zhang C, Chen P. Bispectral index monitoring of the clinical effects of propofol closed-loop target-controlled infusion: Systematic review and meta-analysis of randomized controlled trials. *Medicine (Baltimore)*. 2021;100(4):e23930.
22. Pasin L, Nardelli P, Pintaudi M, et al. Closed-loop delivery systems versus manually controlled administration of total IV anesthesia: a meta-analysis of randomized clinical trials. *Anesth Analg*. 2017;124(2):456-464.
23. Taylor KI, Staunton H, Lipsmeier F, Nobbs D, Lindemann M. Outcome measures based on digital health technology sensor data: data- and patient-centric approaches. *NPJ Digit Med*. 2020;3:97.
24. Sutton R, Barto A. *Reinforcement Learning: An Introduction*. 2nd ed. MIT Press; 2018.
25. García J, Fernández F. A comprehensive survey on safe reinforcement learning. *J Mach Learn Res*. 2015;16(1):1437-1480.
26. Ribba B, Dudal S, Lavé T, Peck RW. Model-informed artificial intelligence: reinforcement learning for precision dosing. *Clin Pharmacol Ther*. 2020;107(4):853-857.
27. Moore BL, Todd EDS, Quasny M, Pyeatt LD. *Intelligent Control of Closed-Loop Sedation in Simulated ICU Patients*. American Association for Artificial Intelligence; 2004.
28. Bräm DS, Parrott N, Hutchinson L, Steiert B. Pharmacokinetic model driven infusion of propofol in children. *Br J Anaesth*. 1991;67(1):41-48.
29. Eleveld DJ, Colin P, Absalom AR, Struys MMRF. Pharmacokinetic-pharmacodynamic model for propofol for broad application in anaesthesia and sedation. *Br J Anaesth*. 2018;120(5):942-959.
30. Lu J, Deng K, Zhang X, Liu G, Guan Y. Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine

learning models in predicting new dosing regimens. *iScience*. 2021;24(7):102804.

31. Bräm DS, Parrott N, Hutchinson L, Steiert B. Introduction of an artificial neural network-based method for concentration-time predictions. *CPT Pharmacometrics Syst Pharmacol*. 2022;11(6):745-754.

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: Ribba B, Bräm DS, Baverel PG, Peck RW. Model enhanced reinforcement learning to enable precision dosing: A theoretical case study with dosing of propofol. *CPT Pharmacometrics Syst Pharmacol*. 2022;11:1497-1510. doi:10.1002/psp4.12858