Received: 21 August 2020 | Revised: 2 December 2020 | Accepted: 10 December 2020

DOI: 10.1002/psp4.12588

ARTICLE

Reinforcement learning and Bayesian data assimilation for model-informed precision dosing in oncology

Corinna Maier[1,2] | **Niklas Hartung**[1] | **Charlotte Kloft**[3] | **Wilhelm Huisinga**[1] | **Jana de Wiljes**[1]

[1]Institute of Mathematics, University of Potsdam, Potsdam, Germany

[2]Graduate Research Training Program PharMetrX: Pharmacometrics & Computational Disease Modelling, Freie Universität Berlin and University of Potsdam, Potsdam, Germany

[3]Department of Clinical Pharmacy and Biochemistry, Institute of Pharmacy, Freie Universität Berlin, Berlin, Germany

Correspondence
Wilhelm Huisinga, Institute of Mathematics, Universität Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam/Golm, Germany.
Email: huisinga@uni-potsdam.de

Funding information
This work was funded by the Graduate Research Training Program PharMetrX: Pharmacometrics & Computational Disease Modelling, Berlin/Potsdam, Germany, Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901 (research & article processing fees).

Abstract

Model-informed precision dosing (MIPD) using therapeutic drug/biomarker monitoring offers the opportunity to significantly improve the efficacy and safety of drug therapies. Current strategies comprise model-informed dosing tables or are based on maximum a posteriori estimates. These approaches, however, lack a quantification of uncertainty and/or consider only part of the available patient-specific information. We propose three novel approaches for MIPD using Bayesian data assimilation (DA) and/or reinforcement learning (RL) to control neutropenia, the major dose-limiting side effect in anticancer chemotherapy. These approaches have the potential to substantially reduce the incidence of life-threatening grade 4 and subtherapeutic grade 0 neutropenia compared with existing approaches. We further show that RL allows to gain further insights by identifying patient factors that drive dose decisions. Due to its flexibility, the proposed combined DA-RL approach can easily be extended to integrate multiple end points or patient-reported outcomes, thereby promising important benefits for future personalized therapies.

INTRODUCTION

Personalized dosing offers the opportunity to improve safety and efficacy of drugs beyond the current practice.[1] This is particularly crucial for drugs that exhibit narrow therapeutic indices relative to the variability between patients. Patient-specific dose adaptations during ongoing treatments, however, are difficult to implement due to the need to integrate multiple sources of information, and labels often only give simplified guidelines for dose adaptations, like dose reductions for severe/life-threatening toxicities.[2,3]

A particularly critical case is cytotoxic anticancer chemotherapy with neutropenia as major dose-limiting toxicity.[4] Patients with severe neutropenia experience a drastic reduction of neutrophil granulocytes and are thus highly susceptible to potentially life-threatening infections. Depending on the lowest neutrophil concentration (nadir), the different grades g of neutropenia range from no neutropenia ($g = 0$)

Wilhelm Huisinga and Jana de Wiljes contributed equally to this work.

to life-threatening ($g = 4$).[5] At the same time, neutropenia serves as a surrogate for efficacy (in terms of median [overall] survival).[6–8] Neutrophil concentrations can therefore be used as a biomarker to guide dosing and therapy management of chemotherapeutic agents that cause neutropenia.[9–11]

In this paper, we consider paclitaxel-induced neutropenia as an illustrative and therapeutically relevant application. Paclitaxel is used as first-line treatment against non-small cell lung cancer in platinum-based combination therapy.[12] The standard dosing of paclitaxel is based on the patient's body surface area (BSA). To individualize treatment, a dosing table based on sex, age, BSA, drug exposure, and toxicity was developed previously[13] and evaluated in a clinical trial (hereafter the "CEPAC-TDM study").[14]

Model-informed precision dosing (MIPD) describes approaches for dose individualization that take into account prior knowledge on the drug-disease-patient system and associated variability (e.g., from a nonlinear mixed effects [NLMEs] analysis) as well as patient-specific therapeutic drug/biomarker monitoring (TDM) data.[15] A popular approach is based on maximum a posteriori (MAP) estimation,[16–18] which infers the individual model parameters of the pharmacokinetic/pharmacodynamic (PK/PD) model. MAP-based outcomes are typically evaluated with respect to a utility function or a target concentration to determine the next dose (MAP-guided dosing).[17,19] The definition of a target concentration or utility function is, however, difficult because in many therapies rather subtherapeutic or toxic ranges are known. For therapeutic ranges, MAP-guided dosing is not readily suited,[20] because only a (potentially biased) point estimate is used, neglecting associated uncertainties.[21] A post hoc uncertainty quantification for MAP-based predictions often relies on a normal approximation located at the MAP estimate, which was previously shown to not necessarily transform accurately into quantities of interest for nonlinear models (e.g., to the nadir concentration[21]).

Reinforcement learning (RL) has been applied to various fields in health care, however, mainly focusing on clinical trial design,[22,23] and only few studies relate to optimal dosing in a PK/PD context.[24,25] In model-based RL, it is learned how to act best in an uncertain environment using model simulations. A key aspect of learning is to make successively use of knowledge already acquired, while also exploring yet unknown sequences of actions. The result is typically a decision tree (or some functional relationship). In other words, the physician's decision is supported via a precalculated, extensive, and detailed look-up table without additional computation during the course of therapy.

Recently, we have shown that Bayesian data assimilation (DA) approaches provide informative clinical decision support, fully exploiting patient-specific information.[21] DA allows for individualized uncertainty quantification, which can be used in a straightforward way (i) to integrate both safety and efficacy aspects into the objective function of finding the optimal dose, or (ii) to accurately compute the probability of being within/outside the target range. However, optimizing across a whole therapy time frame can be hard and potentially too costly for real-time decision support.

In this paper, we demonstrate how DA and RL can be very beneficially exploited to develop new approaches to MIPD. The first approach, referred to as DA-guided dosing, improves existing online MIPD by integrating model uncertainties into the dose selection process. For the second approach (RL-guided dosing) we propose the Monte Carlo tree search (MCTS) in conjunction with the upper confidence bound applied to trees (UCT)[26,27] as sophisticated learning strategy. The third approach combines DA and RL (DA-RL-guided) to make full use of patient TDM

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Current strategies to model-informed precision dosing (MIPD) are either static model-informed dosing tables or maximum a posteriori estimate-based approaches.

WHAT QUESTION DID THIS STUDY ADDRESS?

How could Bayesian data assimilation (DA) and reinforcement learning (RL) methodology be used and combined to advance current approaches towards MIPD?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

We propose more comprehensive approaches to MIPD, which use RL for complex patient state/dose combinations and/or Bayesian DA for individualized uncertainty quantification and propagation to the therapeutic outcome. The combination of the two approaches allows efficient allocation of computational resources and brings together the advantages of the individual approaches. We compare these novel dosing strategies with traditional approaches to control chemotherapy-induced neutropenia.

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

Well-informed and efficient MIPD bears the potential to bring potentially safe and efficacious drugs with narrow therapeutic index and/or high between patient variability to the market and improve therapeutic management in individual patients under-represented in clinical studies.

data and to provide a flexible, interpretable, and extendable framework. We compared the three proposed approaches with current dosing strategies in terms of dosing performance and their ability to provide insights into the factors driving dose selection.

METHODS

We consider a single dose every 3 weeks schedule for paclitaxel-based chemotherapy, usually termed a cycle $c = 1, \cdots, C$, for a total of 6 cycles ($C = 6$). We denote the decision timepoint for the dose of cycle c by T_c, and assume $T_1 = 0$ (therapy start). For dose selection, the physician has different sources of information available, such as the patient's covariates "cov" (sex, age, etc.), the treatment history (drug, dosing regimen, etc.), and TDM data related to PK/PD (drug concentrations, response, toxicity, etc.). Despite these multiple sources of information, it remains a partial and imperfect information problem, as only noisy measurements of few quantities of interest at certain timepoints are available. MIPD aims to provide decision support by linking prior information on the drug-patient-disease system with patient-specific TDM data.

The standard dosing for 3-weekly paclitaxel, as applied in the CEPAC-TDM study arm A, is $200\text{mg}/\text{m}^2\text{BSA}$and a 20% dose reduction if neutropenia grade 4 ($g_c = 4$) was observed.[14] The aforementioned dosing table (termed PK-guided dosing[13]) was evaluated in study arm B, see Section S3 in Appendix S1. For dose selection at cycle start T_c, we chose the patient state:

$$s_{c-1} = (\text{sex}, \text{age}; \text{ANC}_0, g_1, \cdots, g_{c-1}), \tag{1}$$

with $s_0 = (\text{sex}, \text{age}; \text{ANC}_0)$. The covariates sex, age, have previously been identified as important predictors of exposure,[13] and baseline absolute neutrophil counts ANC_0, as a crucial parameter in the drug-effect model.[28,29] We included the neutropenia grades of all previous cycles $g_{1:c-1} = (g_1, \cdots, g_{c-1})$ to account for the observed cumulative behavior of neutropenia.[29,30]

MIPD framework

MIPD builds on prior knowledge from NLME analyses of clinical studies.[21] The structural and observational models are generally given as:

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = f(x(t); \theta, d), \quad x(0) = x_0(\theta) \tag{2}$$

$$h(t) = h(x(t), \theta) \tag{3}$$

with state vector $x = x(t)$ (e.g., neutrophil concentration), parameter values θ (e.g., mean transition time), and rates of change $f(x; \theta, d)$ for given doses d. The initial conditions x_0 are given by the pretreatment levels (e.g., ANC_0). A statistical model links the observables, the quantities that can be measured $h_j(\theta) = h\left(x\left(t_j\right), \theta\right)$ at timepoints t_j to observations $(t_j, y_j)_{j=1,\ldots,n}$ taking into account measurement errors and potential model misspecifications, for example:

$$Y_{j|\Theta=\theta} = h_j(\theta) + \varepsilon_j \tag{4}$$

with $\varepsilon_j \sim_{iid} \mathcal{N}(0, \Sigma)$. In more general terms, $Y_{j|\Theta=\theta} \sim p(\cdot\,|\theta; h_j(\theta), \Sigma)$, with $j = 1, \cdots, n$ independent. The prior distribution for the individual parameters is given by a covariate and statistical model:

$$\Theta \sim p_\Theta\left(\theta^{\text{TV}}(\text{cov}), \Omega\right) \tag{5}$$

with $\theta^{\text{TV}}(\text{cov})$ denoting the typical values, which generally depend on covariates "cov," and Ω the magnitude of the interindividual variability. We used the term "model" to refer to Equations 2–5, and the term "model state of the patient" to refer to a model-based representation of the state of the patient (i.e., a distribution of state-parameter pairs (x, θ) or just a single [reference] state-parameter pair). In the proposed approaches, the model is used to simulate treatment outcomes (in RL called "simulated experience"), or to assimilate TDM data and infer the model state of the patient, or both. To infer the patient state in Equation 1, the grade of neutropenia of the previous cycle g_{c-1} needs to be determined; either directly from the TDM data ($y_{c-1} \mapsto g_{c-1} \mapsto s_{c-1}$) or based on a model simulation of the model state of the patient ($(x, \theta) \mapsto c_{\text{nadir}} \mapsto g_{c-1} \mapsto s_{c-1}$). Because generally measurements are not taken exactly at the time of nadir, the model-predicted nadir may provide an improved state estimate.

We considered three different approaches toward MIPD, see Figure 1:

(i) *Offline approaches* support dose individualization based on precalculated model-informed dosing tables (MIDTs) or dosing decision trees (RL-guided dosing). At the start of therapy, a dose based on the patient's covariates and baseline measurements is recommended. During therapy, the observed TDM data are used to determine a path through the table/tree; whereas the treatment is individualized to the patient (based on a priori uncertainties), the procedure of dose individualization itself does not change (i.e., the tree/table is static). As such, it can be communicated to the physician before the start of therapy.

(ii) *Online approaches* determine dose recommendations based on a model state of the patient and its simulated

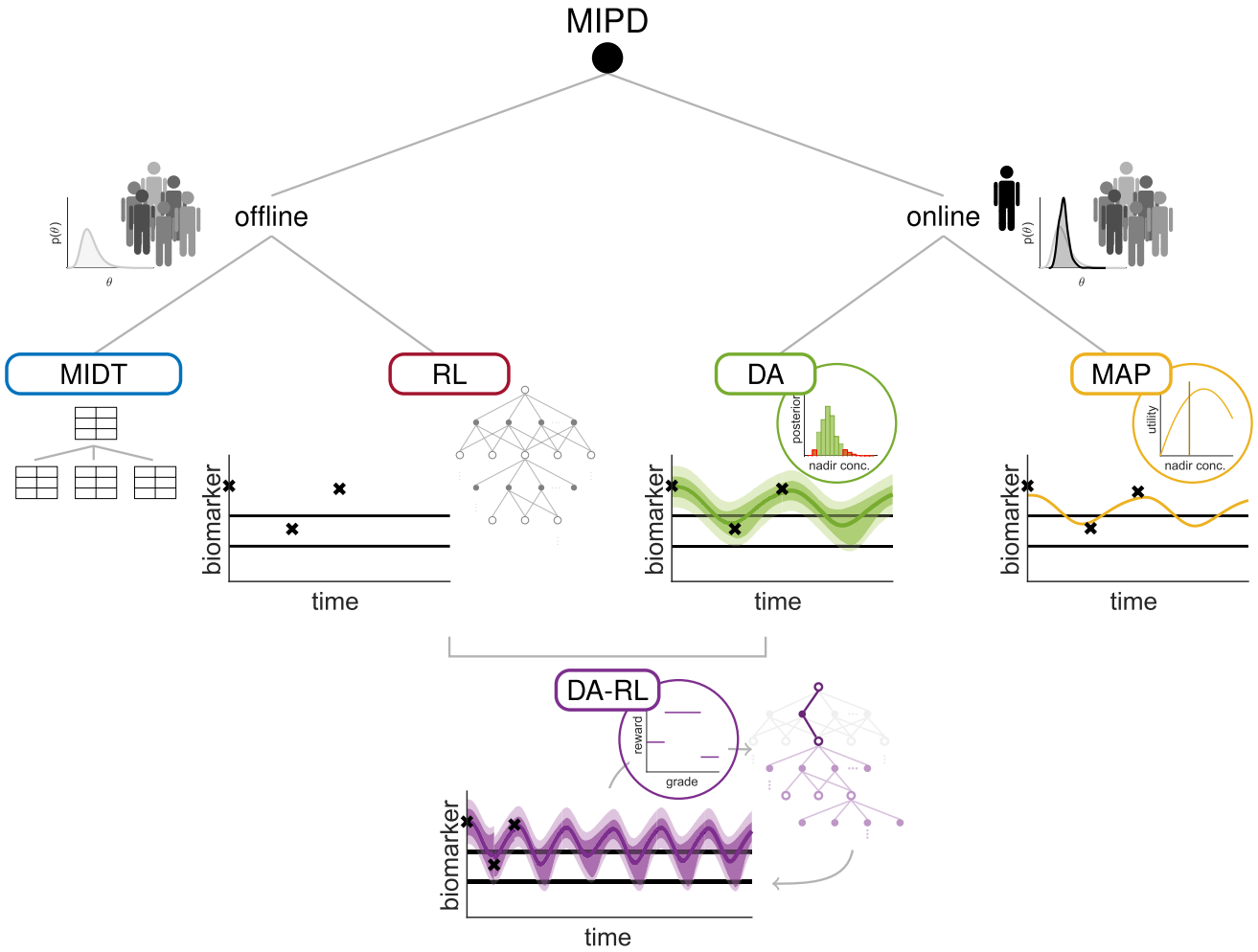


FIGURE 1 Overview of different approaches for model-informed precision dosing (MIPD). The different methods can be categorized according to the time when the computational effort to calculate the optimal dose must be made. Offline approaches calculate optimal doses for all possible covariates and state combinations prior to any treatment, like in precalculated model-informed dosing tables (MIDT) and reinforcement learning (RL). The physician selects the dosing recommendation in the table/tree based on specific patient information (covariates, observations). Although the therapeutic drug/biomarker monitoring (TDM) data (measured biomarker) are used to determine the entry in the table/tree, the table/tree itself is static. Online approaches solve an optimization problem at any decision time point (i.e., when a dose has to be given). They integrate patient-specific TDM data using Bayesian data assimilation (DA) or maximum a posteriori (MAP) estimation. Offline-online approaches allocate computational resources between offline and online. Precalculated dosing decision-trees are individualized during treatment, based on TDM data

outcome. Bayesian DA or MAP estimation assimilate individual TDM data to infer the posterior distribution or MAP point-estimate as model state of the patient, respectively. Although online approaches tailor the model (more precisely, the parameters) to the patient, clinical implementation requires an information technology infrastructure and/or easy-to-use software. Whereas this might constitute a challenging problem that hinders broad application,[31] successful examples of implementation already exist.[32]

(iii) *Offline–Online approaches* combine the advantages of dosing decision trees and an individualized model. The individualized model is used in two ways, to infer the patient state more reliably than sparsely observed TDM data and to individualize the dosing decision tree (using individualized uncertainties, rather than population-based uncertainties).

Key to all approaches is the so-called reward function R (RL terminology), also termed cost or utility function, defined on the set S of patient states:

$$R: S \to \mathbb{R}. \tag{6}$$

Ideally, the reward corresponds to the net utility of beneficial and noxious effects in a patient given the current state.[33] For neutrophil-guided dosing, a reward function was suggested that maps (MAP-based) nadir concentrations to a continuous score[19] or penalizes the deviation from a target nadir concentration ($c_{\text{nadir}} = 1 \cdot 10^9 \text{cells/L}$)[17]; in this study, we used a utility

function but also provide a comparison of the results with the suggested target concentration, see also Section S8.5 in Appendix S1 and Figure S8. The individualized uncertainties quantified via DA allow to consider the probability of being within/outside the target range in the reward function,[21] which is more closely related to clinical reality. For the patient state of Equation 1 used in RL, we also designed the reward function to account for efficacy and toxicity. We chose to penalize the short-term goal (avoiding life-threatening grade 4) more than the long-term goal (increased median [overall] survival) associated with neutropenia grades 1–4[8]:

$$R\left(s_c\right) = \begin{cases} -1 \text{ if } g_c = 0, \\ -1 \text{ if } g_c = 1,2,3, \\ -2 \text{ if } g_c = 4. \end{cases} \tag{7}$$

RL-guided dosing

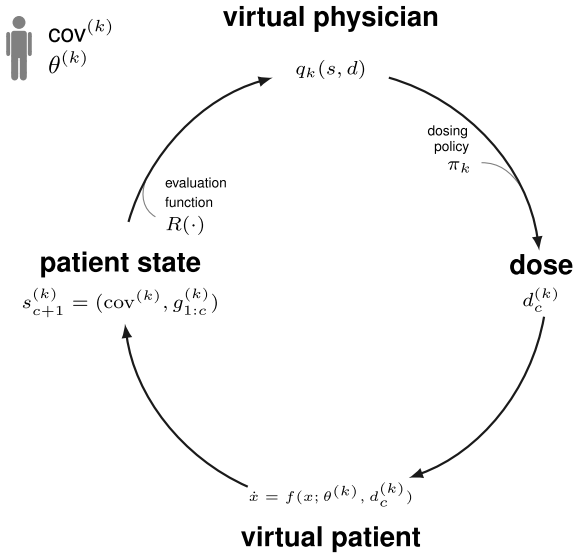
RL problems can be formalized as Markov decision processes, modeling sequential decision making under uncertainty, and are closely related to stochastic optimal control.[34] In RL, the goal of a so-called agent (here, the virtual physician) is to learn a policy (strategy) of how to act (dose) best with respect to optimizing a specific expected long-term return (response), given an uncertain and delayed feedback environment (virtual patient),[35–37] see Figure 2. An introduction to RL in the context of MIPD in descriptive text can be found in ref. 38.

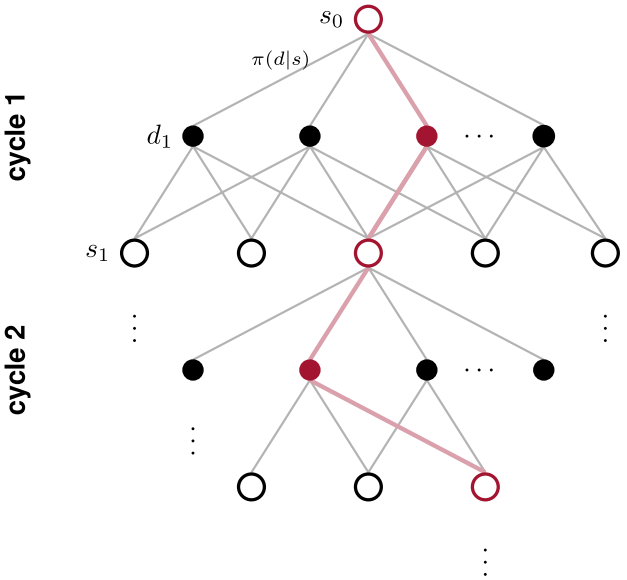
A Markov decision process comprises a sequence including states S_c, actions D_c and rewards R_c with the subscript c referring to time (e.g., treatment cycle). If there is a natural notion of a final time $c = C$ (e.g., therapy of 6 cycles), the sequence is called an episode. Every episode corresponds to a path in the tree of possibilities (Figure 2). Due to unexplained variability between patients (and occasions), transitions between states are characterized by transition probabilities $\mathbb{P}[S_{c+1} = s_{c+1} \,|\, S_c = s_c, D_{c+1} = d_{c+1}]$. The reward is defined via the reward function (i.e., $R_c = R(S_c)$), whereas a so-called dosing policy π models how to choose the next dose:

$$\pi(d|s) = \mathbb{P}[D_{c+1} = d \,|\, S_c = s]. \tag{8}$$

Thus, the policy defines the behavior or strategy of the virtual physician (agent). A dosing policy is evaluated based on the so-called return G_c at time step c, defined as the weighted sum of rewards over the remaining course of therapy:

$$G_c = R_{c+1} + \gamma R_{c+2} + \cdots + \gamma^{C-(c+1)} R_C = \sum_{k=c+1}^{C} \gamma^{k-(c+1)} R_k. \tag{9}$$





sample episode : $s_0^{(k)} \xrightarrow{d_1^{(k)}} (s_1^{(k)}, r_1^{(k)}) \xrightarrow{d_2^{(k)}} (s_2^{(k)}, r_2^{(k)}) \xrightarrow{d_3^{(k)}} \ldots \xrightarrow{d_C^{(k)}} (s_C^{(k)}, r_C^{(k)})$

FIGURE 2 Model-based reinforcement learning (planning). The expected long-term return (action-value function) is estimated based on simulated experience (sample approximation Equation 13). For simulating experience, an ensemble of virtual patients is generated, $k = 1 \cdots, K$ (for all covariate [cov] classes $\mathcal{COV}_l$, $l = 1, \cdots, L$, covariates $\text{cov}^{(k)}$ are sampled within the covariate class and model parameters $\theta^{(k)}$ are sampled from the prior distribution). At the start of each cycle C, a dose $d_c^{(k)}$ is chosen according to the current policy π_k, and the outcome (grade of neutropenia) is predicted based on the model $\dot{x} = f(x;\theta^{(k)}, d_c^{(k)})$ for the sample parameter vector $\theta^{(k)}$ and chosen dose. The updated patient state $s_{c+1}^{(k)}$ is assessed using the reward function R. The sequential dose selections (going through the circle C times [left part]) lead to so-called sample episodes; the entirety of episodes to a tree structure

The discount factor $\gamma \in [0,1]$ balances between short-term ($\gamma \to 0$) and long-term ($\gamma \to 1$) therapeutic goals (see Sections S6 and S8.7 in Appendix S1). Ultimately, the objective is to maximize the expected long-term return:

$$q_\pi(s,d) := \mathbb{E}_\pi\left[G_c | S_c = s, D_{c+1} = d\right], \tag{10}$$

given the current state $S_c = s$ and dose $D_{c+1} = d$ over the space of dosing policies π. The function q_π is called the action-value function.[36] Learning an optimal policy involves maximizing the expected long-term return q_π, which in turn depends on the current estimate of π. Therefore, RL approaches typically involve an iterative process of value estimation and policy improvement.[36]

Model-based RL methods that rely on sampling (sample-based planning) estimate the expected value in Equation 10 via a sample approximation. To simplify the calculations, we have discretized "age" and ANC_0 into covariate classes $\mathcal{COV}_l$, $l = 1, \cdots, L$. For each class $\mathcal{COV}_l$, consider the sample (also called ensemble):

$$\mathcal{E}_{\mathrm{RL}}(\mathcal{COV}_l) := \left\{\left(x_0\left(\theta_c^{(k)}\right), \theta_c^{(k)}, \mathrm{cov}^{(k)}\right)\right\}_{k=1}^{K} \tag{11}$$

with $\mathrm{cov}^{(k)}$ sampled within $\mathcal{COV}_l$ according to the covariate distributions in the CEPAC-TDM study,[14,39] parameter values sampled from $p_\Theta\left(\theta^{\mathrm{TV}}\left(\mathrm{cov}^{(k)}\right), \Omega\right)$ and initial states according to Equation 2. Then, for each $k = 1, \cdots, K$ with K large, a sample episode:

$$s_0^{(k)} \xrightarrow{d_1^{(k)}} \left(s_1^{(k)}, r_1^{(k)}\right) \xrightarrow{d_2^{(k)}} \left(s_2^{(k)}, r_2^{(k)}\right) \xrightarrow{d_3^{(k)}} \cdots \xrightarrow{d_C^{(k)}} \left(s_C^{(k)}, r_C^{(k)}\right) \tag{12}$$

using policy π_k is determined and:

$$q_k(s,d) = \frac{1}{N_k(s,d)} \sum_{k'=1}^{k} \sum_{c=1}^{C} 1_{\left(s_c^{(k')} = s,\, d_{c+1}^{(k')} = d\right)} G_c^{(k')} \tag{13}$$

computed. Here, $N_k(s,d)$ denotes the number of times that dose d was chosen in patient state s among the first k episodes, and $G_c^{(k)} = r_{c+1}^{(k)} + \gamma r_{c+2}^{(k)} + \cdots$. Ideally, $N_k(s,d)$ should be large for each state-dose combination to guarantee a good approximation of the expected return (law of large numbers). This, however, is infeasible for most applications (curse of dimensionality). Thus, one is confronted with the trade-off between exploitation (choosing the doses that are known to give a high return) and exploration (trying new doses that potentially lead to an even higher return). In RL, methods have been developed to cope with this trade-off; we used MCTS in conjunction with UCT as policy in the iterative training process:[26,27,40–42]

$$\pi_{k+1}\left(d_{c+1} | s_c\right) = \begin{cases} 1 \text{ if } d_{c+1} = \arg\max\limits_{d \in D} \mathrm{UCT}_k\left(s_c, d\right) \\ 0 \qquad\qquad \text{else} \end{cases}, \tag{14}$$

with UCT_k defined based on the current sample estimate $q_k(s_c, d)$:

$$\mathrm{UCT}_k\left(s_c, d\right) = \underbrace{q_k\left(s_c, d\right)}_{\text{exploitation}} + \underbrace{\varepsilon_c \frac{\sqrt{N_k\left(s_c\right)}}{N_k\left(s_c, d\right) + 1}}_{\text{exploration}}. \tag{15}$$

It successively expands the search tree (Figure 2) by focusing on promising doses (exploitation, large $q_k(s_c, d)$), while also encouraging exploration of doses that have not yet been tested exhaustively (small $N_k(s_c, d)$relative to the total number of visits $N_k\left(s_c\right) := \sum_{d'} N_k\left(s_c, d'\right)$ to state s_c). The parameter $\varepsilon_c$balances exploration versus exploitation; it depends on the range of possible values of the return and current state of the therapy (cycle c), see Equation S10 in Appendix S1. Finally, we define $\hat{\pi}_{\mathrm{UCT}} = \pi_K$as estimate of the optimal dosing policy in the training setting (learning with virtual patients), and $\hat{q}_{\pi_{\mathrm{UCT}}} = q_K$as an estimate of the associated expected long-term return. In a clinical TDM setting (RL-guided dosing), we finally use $\pi^* = \arg\max \hat{q}_{\pi_{\mathrm{UCT}}}$, i.e., $\varepsilon_c = 0$ (no exploration) in Equation 15. See Section S6.1 in Appendix S1 for details.

DA-guided dosing

Sequential DA approaches have been introduced as more informative alternatives to MAP-based predictions of the therapy outcome, because they provide unbiased predictions of the therapy outcome and a comprehensive uncertainty quantification at the level of the parameters and quantities of interest (e.g., neutropenia grades).[21] The article by Maier et al.,[21] provides a more thorough introduction to DA in the context of pharmacometrics, including the interpretation of domain-specific terminology. The individualized uncertainty in the model state of the patient is inferred and propagated to the predicted therapy time course, allowing to predict the probability of possible outcomes. For this, the uncertainty in the individual model parameters is sequentially updated via Bayes' formula, that is:

$$p\left(\theta | y_{1:c}\right) \propto p\left(y_c | \theta\right) \cdot p\left(\theta | y_{1:c-1}\right), \tag{16}$$

where $y_{1:c} = (y_1, \ldots, y_c)^T$ denotes the TDM data up to and including cycle c, and $y_c = (y_{c1}, \ldots, y_{cn_c})^T$ the measurements taken in cycle c. Because the posterior distribution $p(\theta | y_{1:c})$ generally cannot be determined analytically, DA approaches approximate it by an ensemble of so-called particles (a sample approximation):

$$\mathcal{E}_{1:c} := \left\{\left(x_{1:c}^{(m)}, \theta_c^m, w_c^{(m)}\right)\right\}_{m=1}^{M}. \tag{17}$$

In our context, a particle represents a potential model state of the patient (for the specific patient cov) with a weighting factor $w_c^{(m)}$ characterizing how probable the state is (given prior knowledge and TDM data up to C). As more TDM data is gathered, the Bayesian updates reduce the uncertainty in the model parameters and consequently in the therapeutic outcome, see Figure 3 (DA part, reduced width of credible intervals/prediction intervals) and Section S5 in Appendix S1. Because subtherapeutic as well as toxic ranges (i.e., very low or high drug/biomarker concentrations), are described by the tails of the posterior distribution, the uncertainties provide crucial additional information compared to the mode (MAP estimate) for dose selection.

We chose the optimal dose to be the dose that minimizes the weighted risk of being outside the target range (i.e., the a posteriori probability of $g_c = 0$ or $g_c = 4$):

$$d_{c+1}^* = \arg\min_{d \in \mathcal{D}} \lambda_0 \sum_{m=1}^{M} w_c^{(m)} \mathbf{1}_{\{g(\theta_c^{(m)}, d)=0\}} + \lambda_4 \sum_{m=1}^{M} w_c^{(m)} \mathbf{1}_{\{g(\theta_c^{(m)}, d)=4\}} \tag{18}$$

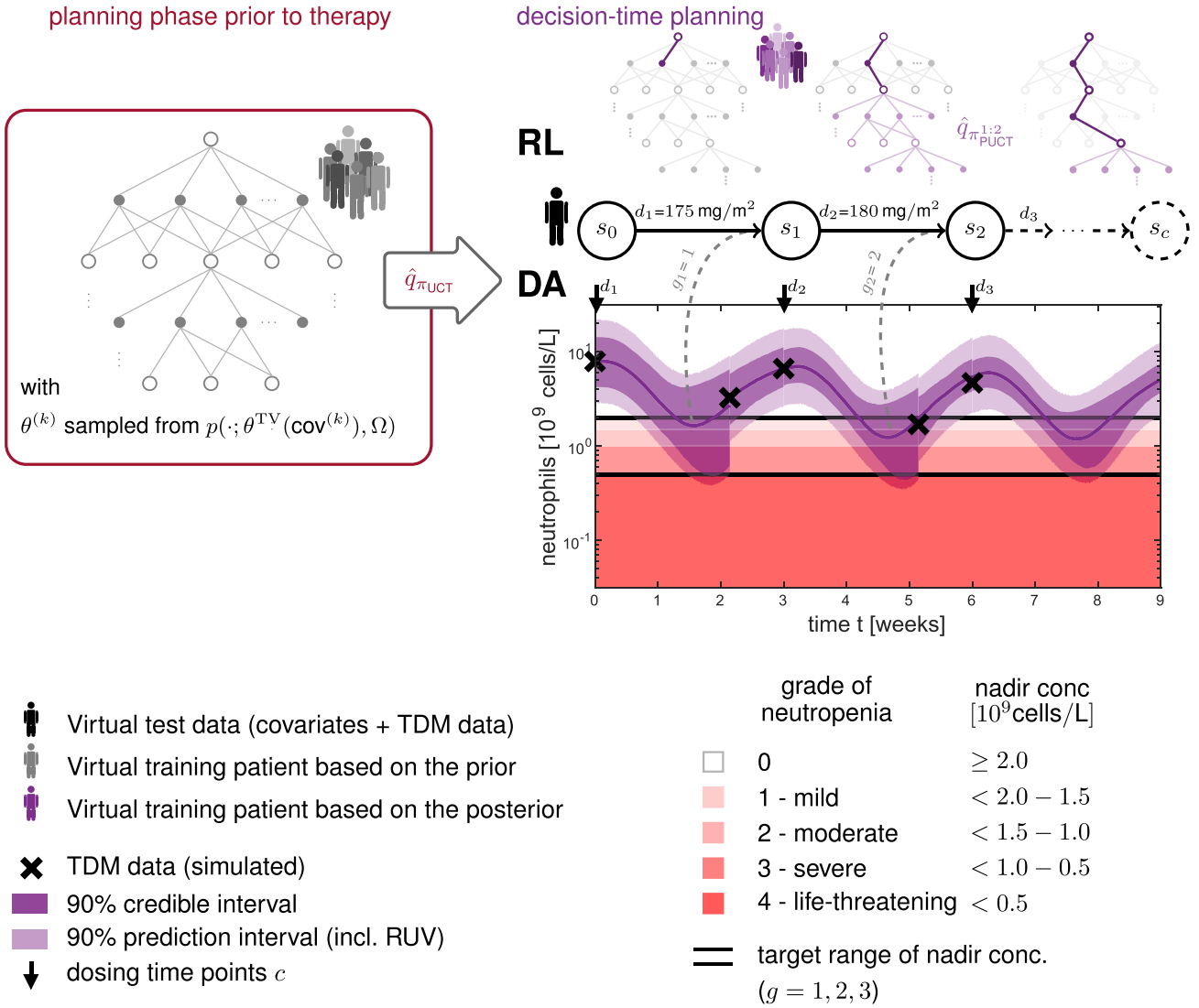


FIGURE 3 The interplay of data assimilation (DA) and reinforcement learning (RL). In the planning phase prior to therapy, the expected long-term return $q_{\pi_0} := \hat{q}_{\pi_{\mathrm{UCT}}}$ is estimated in Monte Carlo Tree Search (MCTS) with upper confidence bound applied to trees (UCT) using an ensemble of covariates $(\mathrm{cov})^{(k)}$ and parameter values $\theta^{(k)}: p(\,|\,\theta^{\mathrm{TV}}\,(\mathrm{cov}^{(k)})\,, \Omega)$. The first dose is selected based on $q_{\pi_0} := \hat{q}_{\pi_{\mathrm{UCT}}}$ for the patient specific covariate class. The DA algorithm initializes a particle ensemble given the patient's covariates. The ensemble is propagated forward continuously in time, and observed patient therapeutic drug/biomarker monitoring (TDM) data (black crosses) is assimilated when it becomes available. This results in updated uncertainty, visible as "cuts" in the credible/prediction intervals. In contrast, the RL state evolves in discrete time steps C according to the decision timepoints and only considers selected features/summaries of the model state of the patient (e.g., smoothed posterior expectation of nadir concentrations translated into neutropenia grades). At each decision timepoint, the posterior model state of the patient is used to refine the prior computed $\hat{q}_{\pi_{\mathrm{UCT}}}$ (grey tree) for future reachable states (light purple tree). This individualizes the tree based on individualized uncertainties ($\mathcal{E}_{1:c}$)

with $g(\theta_c^{(m)}, d)$ denoting the predicted neutropenia grade by forward simulation of the m-th particle for dose d. We penalized grade 4 more severely than grade 0, i.e., $\lambda_4 = 2/3$ and $\lambda_0 = 1/3$, similarly as in Equation 7.

The integration of an ensemble of particles into the optimization problem, instead of a point estimate (as in MAP-guided dosing), increases the computational effort and complexity of the problem.

If time or computing power is limited, approximations have to be used (e.g., by solving only for the next cycle dose rather than all remaining cycles at the cost of neglecting long-term effects). Alternatively, the number of particles M could be reduced (we used both approximations in this study); see also Section S8.6 in Appendix S1. The DA optimization problem is stated in the space of actions (doses), whereas RL optimizes in the space of states by estimating the expected long-term return as an intermediate step (Equation 13) thereby promising efficient solutions to the sequential decision making problem under uncertainty.[36]

DA-RL-guided dosing

The particle-based DA scheme and the model-based RL scheme address the problem of personalized dosing from different angles. A combined DA-RL approach therefore offers several advantages by integrating individualized uncertainties provided by DA within RL, see Figure 3. First, instead of the observed grade (e.g., measured neutrophil concentration on a given day, translated into the neutropenia grade), we may use the smoothed posterior expectation of the quantity of interest (e.g., predicted nadir concentration), see Section S7 in Appendix S1. This reduces the impact of measurement noise and the dependence on the sampling day. Second, for model simulations within the RL scheme, we can sample from the posterior $p(\theta \mid y_{1:c})$ represented by the ensemble $\mathcal{E}_{1:c}$ (i.e., from individualized uncertainties, instead of the prior $p(\theta)$, i.e., population-based uncertainties). During the course of the treatment, the ensemble of potential model states of the patient is continuously updated when new patient-specific data are obtained (see Equation 16). This allows to individualize the expected long-term return during treatment as new patient data is observed, see Figure 3 (i.e., the dosing decision tree in RL is updated prior to the next dosing decision).

Because the refinement as well as the DA part has to run in real-time (online), it has to be performed efficiently. We do not need to take all possible state combinations into account, but only those that are still relevant for the remaining part of the therapy. This reduces the computational effort, in particular for later cycles. The proposed DA-RL approach results in a sequence of estimated optimal dosing policies $\hat{\pi}^1, \hat{\pi}^{1:2}, \cdots$ with $\hat{\pi}^{1:c}$ denoting the estimated optimal dosing policy based on TDM data $y_{1:c}$ (i.e., based on $\mathcal{E}_{1:c}$). In addition, we do not need to estimate the individualized action-value function from scratch, but can exploit $q_{\pi_0} := \hat{q}_{\pi_{\mathrm{UCT}}}$ as a prior determined by the RL scheme prior to any TDM data (see paragraph following Equation 15). In predictor + UCT (PUCT[27,43]), the exploitation versus exploration parameter ε_c in Equation 15 is modified to prioritize doses with high a priori expected long-term return:

Algorithm 1 DA-RL guided dosing

Sample particles to get ensemble $\mathcal{E}_0$ from the prior $p_\Theta\left(\cdot; \theta^{\mathrm{TV}}(\mathrm{cov}), \Omega\right)$
Get s_0 based on covariates and baseline measurement
Choose optimal dose $d_1^* = \arg\max_{d \in \mathcal{D}} \hat{q}_{\pi_{\mathrm{UCT}}}(s_0, d)$
for $c = 1 : C$ **do**
  $\mathcal{E}_{1:c} \leftarrow$ update ensemble $\mathcal{E}_{1:c-1}$ by assimilating data y_c (DA part)
  $s_c \leftarrow$ posterior expectation under ensemble $\mathcal{E}_{1:c}$
  $\hat{q}_{\pi^{1:c}_{\mathrm{PUCT}}} \leftarrow$ MCTS with PUCT using the ensemble $\mathcal{E}_{1:c}$
  Choose optimal dose $d_{c+1}^* = \arg\max_{d \in \mathcal{D}} \hat{q}_{\pi^{1:c}_{\mathrm{PUCT}}}(s_c, d)$
end for

FIGURE 4 Pseudo code of DA-RL-guided dosing. At therapy start a particle ensemble $\mathcal{E}_0$ for the sequential data assimilation (DA) approach is sampled from the prior parameter distribution given the patient's covariates. Then for the initial state s_0 (pretreatment) the first dose is selected according to the prior expected long-term return $q_{\pi_0} := \hat{q}_{\pi_{\mathrm{UCT}}}$ calculated beforehand in the prior planning phase (Monte Carlo Tree Search [MCTS] with upper confidence bound applied to trees [UCT]). The selected dose is given to the patient and patient-specific therapeutic drug/biomarker monitoring (TDM) data y_c is collected within cycle c. The TDM data is assimilated via a sequential DA approach (particle filter and smoother) creating a posterior particle ensemble $\mathcal{E}_{1:c}$. For subsequent dose decisions ($c = 1, \cdots, C$). The new patient state is inferred using $\mathcal{E}_{1:c}$ (e.g., smoothed posterior expectation of the nadir concentration translated into the neutropenia grade of the cycle). Then an MCTS is started from the current patient state using $\mathcal{E}_{1:c}$ in the model simulations. Within the tree search we use the PUCT algorithm with prioritized exploration based on $\hat{q}_{\pi_{\mathrm{UCT}}}$. PUCT, predictor upper confidence bound applied to trees; RL, reinforcement learning

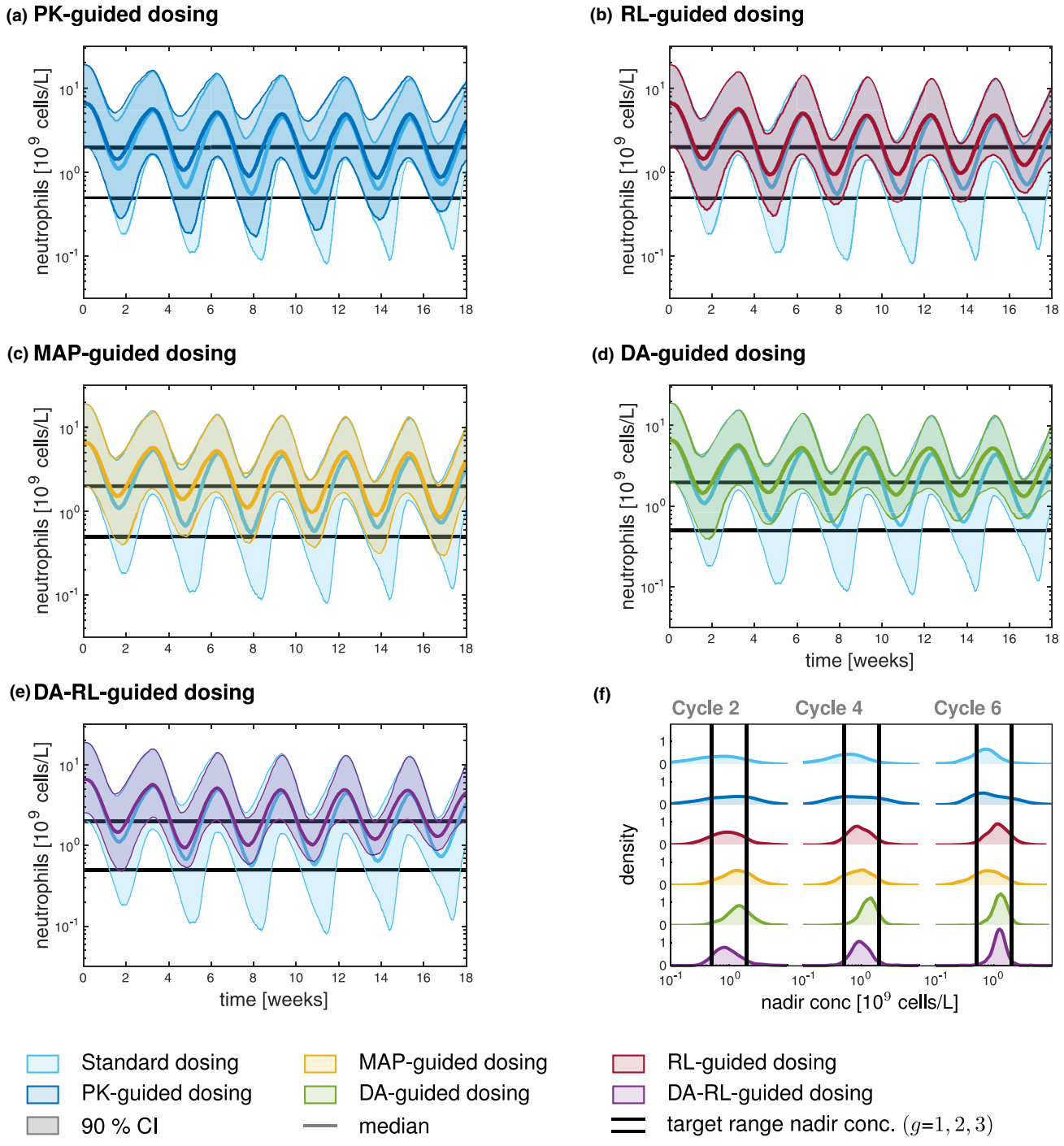


FIGURE 5 Comparison of different dosing policies for paclitaxel dosing. Comparison of the 90% confidence intervals (CIs) and median of the neutrophil concentration for the test virtual population ($N = 1000$) using (a) pharmacokinetic (PK)-guided dosing, (b) reinforcement learning (RL)-guided dosing, (c) maximum a posteriori (MAP)-guided dosing, (d) data assimilation (DA)-guided dosing, and (e) DA-RL-guided dosing, each in comparison to the standard dosing (body surface area [BSA]-based dosing). PK-guided dosing is the only approach that also takes into account exposure data ($T_{C_{\text{drug}} \geq 0.05 \mu\text{mol/L}}$). (f) Comparison of the distributions of model-predicted nadir concentrations (smooth by kernel density estimation) for the test virtual population ($N = 1000$)

$$U_k(s_c, d) = \underbrace{q_k^{1:c}(s_c, d)}_{\text{exploitation}} + \varepsilon_c \cdot \underbrace{\frac{\exp\left(\hat{q}_{\pi_{UCT}}(s, d)\right)}{\sum_{d'} \exp\left(\hat{q}_{\pi_{UCT}}(s, d')\right)}}_{\text{priortizing}} \underbrace{\frac{\sqrt{N_k(s_c)}}{N_k(s_c, d) + 1}}_{\text{exploration}}. \tag{19}$$

Finally, we define $\hat{\pi}^{1:c}_{\text{PUCT}} = \pi^{1:c}_K$ based on $\mathcal{E}_{1:c}$ as an estimate of the optimal individualized dosing policy in the training setting (using Equations 14 and 19), and $\hat{q}_{\pi^c_{\text{PUCT}}} = q_K$ as an estimate of the associated expected long term return based on $\mathcal{E}_{1:c}$. For individualized dose recommendations in a clinical TDM setting, we again use $\pi^* = \arg\max \hat{q}_{\pi^{1:c}_{\text{PUCT}}}$ (i.e., $\varepsilon_c = 0$ in

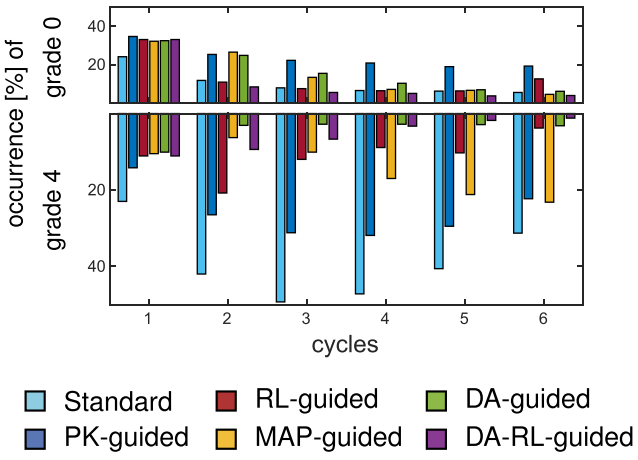


FIGURE 6 Occurrence of grade 0 and grade 4 for the different dosing policies. The percentage is based on a test virtual population ($N = 1000$) and six cycles (inferred from the model predicted nadir concentration). Additional analyses are provided in the supplement, Figure S22. DA, data assimilation; MAP, maximum a posteriori; PK, pharmacokinetic; RL, reinforcement learning

Equation 19; see Figure 3 and 4, and Section S7 in Appendix S1).

RESULTS

Novel individualized dosing strategies decreased the occurrence of grade 4 and grade 0 neutropenia compared with existing approaches

We compared our proposed approaches with existing approaches for MIPD based on simulated TDM data in paclitaxel-based chemotherapy. The design was chosen to correspond to the CEPAC-TDM study[14]: neutrophil counts at days 0 and 15 of each cycle were simulated for virtual patients using a PK/PD model for paclitaxel-induced cumulative neutropenia (Figure S1).[29] We focused only on paclitaxel dosing; we did not take into account drop-outs, dose reductions due to nonhematological toxicities, adherence, and comedication. Occurrence of grade 4 neutropenia, therefore, differed between our simplified simulation study and the clinical study (as might be expected), see Section S8.2 in Appendix S1. This should be taken into account when interpreting the results. To obtain meaningful statistics, all analyses were repeated 1000 times with covariates sampled from the observed covariate ranges in the CEPAC-TDM study. Detailed discussions and further analyses are provided in Sections S2 and S8 in Appendix S1. The MATLAB code is available under (https://doi.org/10.5281/zenodo.3967011).

Figure 5 shows the predicted neutrophil concentrations—median and 90% confidence interval (CI)—over 6 cycles of 3 weeks each. A successful neutrophil-guided dosing should result in nadir concentrations within the target range (grades 1–3, between black horizontal lines). In all cycles, PK-guided dosing prevented the nadir concentrations (90% CI) to drop as low as for the standard dosing (Figure 5a). However, PK-guided dosing also increased the occurrence of grade 0 (Figure 6).

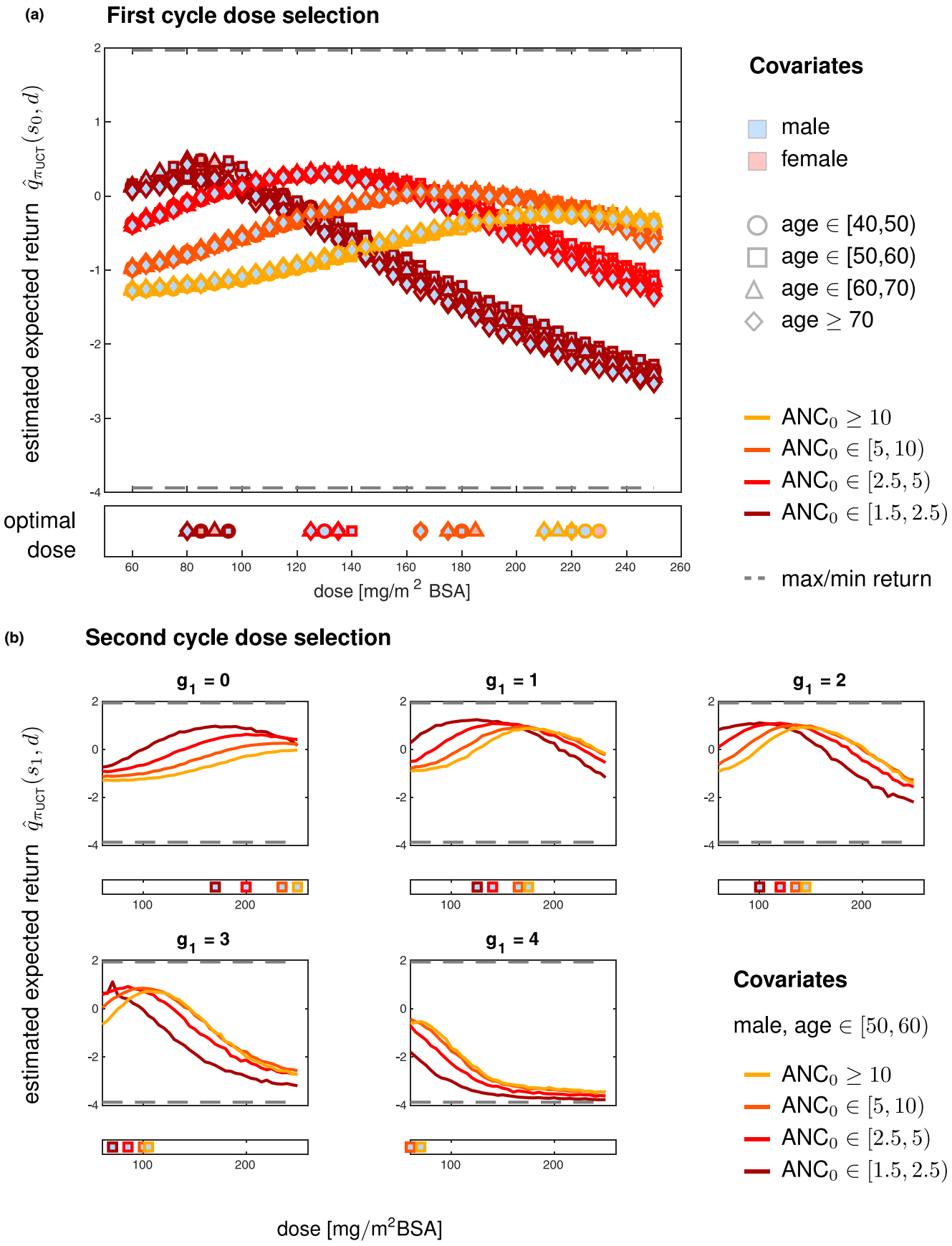
RL-guided dosing controlled the neutrophil concentration well across the cycles (Figure 5b) and the distribution of nadir concentrations over the whole population was increasingly concentrated within the target range (Figure 5f). The occurrence of grade 0 and 4 neutropenia was substantially reduced compared to standard and PK-guided dosing (Figure 6).

For MAP-guided dosing, the occurrence of grade 4 neutropenia increased over the cycles (Figure 6), showing the typical cumulative trend of neutropenia,[29] despite inclusion of TDM data. In contrast, DA steadily guided nadir concentrations into the target range (Figure 5d,f), thereby substantially decreasing the variance (i.e., the variability in outcome). The occurrence of grade 0 and 4 was reduced considerably in later cycles (Figure 6), suggesting that individualized uncertainty quantification played a crucial role in reducing the variability in outcome.

Integrating individualized uncertainties and considering the model state of the patient in the RL approach (DA-RL-guided dosing) also moved nadir concentrations into the target range and clearly decreased the variance (Figure 5b,f). The slight differences between DA and DA-RL (Figure 6) might be related to the difference in weighting grade 0 and 4 in the respective reward functions (Equation 18 vs. Equation 7). For additional comparisons, see Figure S22.

In summary, individualized uncertainties as in DA-guided and DA-RL-guided dosing seemed to be crucial in bringing nadir concentrations into the target range and reducing the variability of the outcome, thus achieving the goal of therapy

FIGURE 7 Expected long-term return across the dose range for dose selection. (a) across the considered covariate combinations for the dose selection in cycle 1. The symbols plotted below the x-axis show the optimal dose for the corresponding covariate class (i.e., the arg max of the plotted line). (b) For fixed sex and age class (here, men between 50 and 60 years) with different pretreatment neutrophil values ANC_0 and observed neutropenia grades in cycle 1 (i.e., g_1). The optimal dose for the second cycle depends on the neutropenia grade of the previous cycle and the pretreatment neutrophil count ANC_0 in (10^9cells/L). The grey dashed line shows the maximum and minimum possible return from the first cycle (a) and the second cycle (b) onwards, with $\gamma = 0.5$. The covariate classes were chosen based on the CEPAC-TDM study population: inclusion criteria were $ANC_0 > \cdot 1.5 \cdot 10^9$cells/L; the typical baseline count for men was $ANC_0 = 6.48 \cdot 1.5 \cdot 10^9$cells/L (arm B). The median age was 63 years ranging from 51 to 74 years (5th and 95th percentile of the population in arm B), see refs. 14,38. ANC, absolute neutrophil count; BSA, body surface area.


(a)
First cycle dose selection
estimated expected return $\hat{q}_{\pi_{UCT}}(s_0, d)$
optimal dose
dose [mg/m^2 BSA]
Covariates
male
female
age ∈ [40,50)
age ∈ [50,60)
age ∈ [60,70)
age ≥ 70
ANC$_0$ ≥ 10
ANC$_0$ ∈ [5, 10)
ANC$_0$ ∈ [2.5, 5)
ANC$_0$ ∈ [1.5, 2.5)
max/min return
(b)
Second cycle dose selection
g_1 = 0
g_1 = 1
g_1 = 2
g_1 = 3
g_1 = 4
estimated expected return $\hat{q}_{\pi_{UCT}}(s_1, d)$
Covariates
male, age ∈ [50, 60)
ANC$_0$ ≥ 10
ANC$_0$ ∈ [5, 10)
ANC$_0$ ∈ [2.5, 5)
ANC$_0$ ∈ [1.5, 2.5)
dose [mg/m^{2}BSA]

individualization. For this specific example, both approaches showed comparable results, but DA-RL has the greater potential for long-term optimization in a delayed feedback environment as well as integrating multiple end points.

Identification of relevant covariates via investigating the expected long-term return in RL

A key object in RL is the expected long-term return or action-value function $q_\pi(s,d)$ (see Equation 10). We demonstrate that it contains important information to identify relevant covariates to individualize dosing.

Figure 7a shows the estimated action-value function for RL-guided dosing stratified for the covariates, sex, age, and baseline neutrophil counts ANC_0 (covariate classes are shown in the legend) for the first cycle dose selection. ANC_0 was found to be by far the most important characteristic for the RL-based dose selection at therapy start. Differences in age and sex played only minor roles. For comparison, the first cycle of dose selection in the PK-guided algorithm is only based on sex and age. The steepness of the curves gives an idea about the robustness of the dose selection.

For the second dose selection, the grade of neutropenia in the first cycle (g_1) has the largest impact, whereas larger ANC_0 led to larger optimal doses (Figure 7b). To illustrate the dose selection in RL, we extracted a similar decision tree to the one developed by Joerger et al.,[13] see Figure S13.

DISCUSSION

We present three promising MIPD approaches using DA and/or RL that substantially reduced the number of (virtual) patients in life-threatening grade 4 and grade 0, a surrogate marker for efficacy of the anticancer treatment.

RL-guided dosing in oncology has been proposed before,[25] however, only considering the mean tumor diameter. Because only a marker for efficacy was considered, this led to a one-sided dosing scheme and resulted in very high optimal doses. The authors therefore introduced action-derived rewards (i.e., penalties on high doses). In contrast, neutrophil-guided dosing considers toxicity and efficacy (link to median survival) simultaneously. Ideally, dosing decisions should also include other adverse effects (e.g., peripheral neuropathy), tumor response, or long-term outcomes (e.g., overall or progression-free survival), and other concomitant medication (anticancer combination agents; e.g., carboplatin and supportive medication; e.g., granulocyte colony stimulating factor and other patient-specific comedications). Notably, RL easily extends to multiple adverse/beneficial effects and comedication, and is especially suited for time-delayed feedback environments,[23,35] as typical in many diseases. Unlike current less complex MIDT, the decision tree of RL is not straightforward to navigate or remember, therefore, an application in clinics would require the development of easy-to-use software or dashboards, as, for example, for infliximab.[32]

So far, RL approaches in health care are limited to rather simple exploration strategies (so-called ϵ-greedy approaches) with one-time step ahead approximations of the look-up table (Q-learning).[23] Using MCTS with UCT, we used an RL framework that exploits the possibility to simulate until the end of therapy and evaluate the return. Consequently, it requires less approximations as temporal difference approaches (e.g., Q-learning, used in ref. 25) that avoid computation of the return via a decomposition (Bellman equation). Moreover, exploration via UCT allows to systematically sample from the dose range (as opposed to an ϵ-greedy strategy) and allows to include additional information (e.g., uncertainties or prior information [as in PUCT]). This becomes key when combined with direct RL based on real-world patient data, see for example,[44,45] which would allow to compensate for a potential model bias. At the end of a patient's therapy, the observed return can be evaluated and used to update the expected return $\hat{q}_{\hat{\pi}}$. This update would even be possible if the physician did not follow the dose recommendation (off-policy learning) and could be implemented across clinics, as it could be done locally without exchanging patient data. Thus, the presented approach builds a basis for continuous learning postapproval, which has the potential to substantially improve patient care, including patient subgroups under-represented in clinical studies.

Overall, we have shown that DA and RL techniques can be seamlessly integrated and combined with existing NLME and data analysis frameworks for a more holistic approach to MIPD. Our study demonstrates that incorporation of individualized uncertainties (as in DA) is favorable over state-of-the-art online algorithms such as MAP-guided dosing. The integrated DA-RL framework allows not only to consider prior knowledge from clinical studies but also to improve and individualize the model and the dosing policy simultaneously during the course of treatment by integrating patient-specific TDM data. Thus, the combination provides an efficient and meaningful alternative to solely DA-guided dosing, as it allocates computational resources between online and offline and the RL part provides an additional layer of learning to the model (in form of the expected long-term return) that can be used to gain deeper insights into important covariates for the dose selection. Therefore, showing that RL approaches can be well interpreted in clinically relevant terms (e.g., highlighting the role of ANC_0 values).

Well-informed and efficient MIPD bears huge potential in drug development as well as in clinical practice as it could: (1) increase response rates in clinical studies, (2) facilitate recruitment by relaxing exclusion criteria, and (3) enable continuous learning postapproval and thus improve treatment outcomes in the long term.

ACKNOWLEDGMENTS

C.M. kindly acknowledges financial support from the Graduate Research Training Program PharMetrX: Pharmacometrics & Computational Disease Modelling, Berlin/Potsdam, Germany. This research and publication of this paper has been funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 – 318763901. Fruitful discussions with Sven Mensing (AbbVie, Germany), Alexandra Carpentier (Otto-von-Guericke-Universitaet Magdeburg), and Sebastian Reich (University of Potsdam, University of Reading) are kindly acknowledged. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

Charlotte Kloft and Wilhelm Huisinga report research grants from an industry consortium (AbbVie Deutschland GmbH & Co. KG, AstraZeneca, Boehringer Ingelheim Pharma GmbH & Co. KG, Grünenthal GmbH, F. Hoffmann-La Roche Ltd., Merck KGaA and Sanofi) for the PharMetrX program. In addition, Charlotte Kloft reports research grants from the Innovative Medicines Initiative-Joint Undertaking ("DDMoRe"), from H2020-EU.3.1.3 ("FAIR") and Diurnal Ltd. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

C.M., N.H., C.K., W.H., and J.dW wrote the manuscript. C.M., N.H., C.K., W.H., and J.dW. designed the research. C.M. performed the research. C.M., N.H., C.K., W.H., and J.dW. analyzed the data.

REFERENCES

1. Peck RW. The right dose for every patient: a key step for precision medicine. *Nat Rev Drug Discov*. 2016;15(3):145–146.
2. de Jonge ME, Huitema ADR, Schellens JHM, Rodenhuis S, Beijnen JH. Individualised Cancer Chemotherapy: Strategies and Performance of Prospective Studies on Therapeutic Drug Monitoring with Dose Adaptation. *Clinical Pharmacokinetics*. 2005;44:147–173. http://dx.doi.org/10.2165/00003088-200544020-00002
3. Darwich AS, Ogungbenro K, Vinks AA, et al. Why has model-informed precision dosing not yet become common clinical reality? Lessons from the past and a roadmap for the future. *Clin Pharmacol Ther*. 2017;101:646–656.
4. Crawford J, Dale DC, Lyman GH. Chemotherapy-induced neutropenia. *Cancer*. 2004;100:228–237.
5. National Cancer Institute. Common terminology criteria for adverse events (CTCAE) version 4.03. Bethesda, Maryland. 1—194. 2010.
6. Cameron DA, Massie C, Kerr G, Leonard RC. Moderate neutropenia with adjuvant CMF confers improved survival in early breast cancer. *Br J Cancer*. 2003;89:1837–1842.
7. Di Maio M, Gridelli C, Gallo C, et al. Chemotherapy-induced neutropenia and treatment efficacy in advanced non-small-cell lung cancer: a pooled analysis of three randomised trials. *Lancet Oncol*. 2005;6:669–677.
8. Di Maio M, Gridelli C, Gallo C, Perrone F. Chemotherapy-induced neutropenia: a useful predictor of treatment efficacy? *Nat Clin Pract Oncol*. 2006;3:114–115.
9. Wallin JE, Friberg LE, Karlsson MO. A tool for neutrophil guided dose adaptation in chemotherapy. *Comput Methods Programs Biomed*. 2009;93:283–291.
10. Hansson EK, Wallin JE, Lindman H, Sandström M, Karlsson MO, Friberg LE. Limited inter-occasion variability in relation to inter-individual variability in chemotherapy-induced myelosuppression. *Cancer Chemother Pharmacol*. 2010;65:839–848.
11. Netterberg I, Nielsen EI, Friberg LE, Karlsson MO. Model-based prediction of myelosuppression and recovery based on frequent neutrophil monitoring. *Cancer Chemother Pharmacol*. 2017;80:343–353.
12. Peters S, Adjei AA, Gridelli C, et al. Metastatic non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2012;23:vii56–vii64.
13. Joerger M, Kraff S, Huitema ADR, et al. Evaluation of a pharmacology-driven dosing algorithm of 3-weekly paclitaxel using therapeutic drug monitoring: a pharmacokinetic-pharmacodynamic simulation study. *Clin Pharmacokinet*. 2012;51:607–617.
14. Joerger M, von Pawel J, Kraff S, et al. Open-label, randomized study of individualized, pharmacokinetically (PK)-guided dosing of paclitaxel combined with carboplatin or cisplatin in patients with advanced non-small-cell lung cancer (NSCLC). *Ann Oncol*. 2016;27:1895–1902.
15. Keizer RJ, Heine R, Frymoyer A, Lesko LJ, Mangat R, Goswami S. Model-informed precision dosing at the bedside: scientific challenges and opportunities. *CPT Pharmacometrics Syst Pharmacol*. 2018;7:785–787.
16. Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. *Clin Pharmacol Ther*. 1979;26:294–305.
17. Wallin JE, Friberg LE, Karlsson MO. Model-based neutrophil-guided dose adaptation in chemotherapy: evaluation of predicted outcome with different types and amounts of information. *Basic Clin Pharmacol Toxicol*. 2009;106:234–242.
18. Bleyzac N, Souillet G, Magron P, et al. Improved clinical outcome of paediatric bone marrow recipients using a test dose and Bayesian pharmacokinetic individualization of busulfan dosage regimens. *Bone Marrow Transplant*. 2001;28:743–751.
19. Wallin JE, Friberg LE, Karlsson MO. Model based neutrophil guided dose adaptation in chemotherapy; evaluation of predicted outcome with different type and amount of information. In Page Meet. 2009.
20. Holford N. Pharmacodynamic principles and target concentration intervention. *Transl Clin Pharmacol*. 2018;26:150–154.
21. Maier C, Hartung N, Wiljes J, Kloft C, Huisinga W. Bayesian data assimilation to support informed decision making in individualized chemotherapy. *CPT Pharmacometrics Syst Pharmacol*. 2020;9:153–164.
22. Zhao Y, Zeng D, Socinski MA, Kosorok MR. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*. 2011;67:1422–1433.
23. Yu C, Liu J, Nemati S. Reinforcement learning in healthcare: a survey. *arXiv*. 2019. https://www.arxiv.org/abs/1908.08796.
24. Escandell-Montero P, Chermisi M, Martínez-Martínez JM, et al. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*. 2014;62:47–60. http://dx.doi.org/10.1016/j.artmed.2014.07.004

25. Yauney G, Shah P. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. *Proc Mach Learn Res*. 2018;85:1-49.
26. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484–489.
27. Rosin CD. Multi-armed bandits with episode context. *Ann Math Artif Intell*. 2011;61:203–230.
28. Friberg LE, Henningsson A, Maas H, Nguyen L, Karlsson MO. Model of chemotherapy-induced myelosuppression with parameter consistency across drugs. *J Clin Oncol*. 2002;20:4713–4721.
29. Henrich A, Joerger M, Kraff S, et al. Semimechanistic bone marrow exhaustion pharmacokinetic/pharmacodynamic model for chemotherapy-induced cumulative neutropenia. *J Pharmacol Exp Ther*. 2017;362:347–358.
30. Huizing MT, Giaccone G, van Warmerdam LJ, et al. Pharmacokinetics of paclitaxel and carboplatin in a dose-escalating and dose-sequencing study in patients with non-small-cell lung cancer. *J Clin Oncol*. 1997;15:317–329.
31. Keizer RJ, Dvergsten E, Kolacevski A, et al. Get real: integration of real-world data to improve patient care. *Clin Pharmacol Ther*. 2020;107:722–725.
32. Dubinsky M, Phan BL, Singh N, et al. Pharmacokinetic dashboard-recommended dosing is different than standard of care dosing in infliximab-treated pediatric IBD patients. *AAPS J*. 2017;19:215–222.
33. Sheiner LB. Learning versus confirming in clinical drug development. *Clin Pharmacol Ther*. 1997;61:275–291.
34. Bertsekas DP. *Reinforcement Learning and Optimal Control*. Belmont, MA: Athena Scientific; 2019.
35. Zhavoronkov A, Vanhaelen Q, Oprea TI. Will artificial intelligence for drug discovery impact clinical pharmacology? *Clin Pharmacol Ther*. 2020;107:780–785.
36. Sutton RS, Barto AG. *Reinforcement learning an introduction*, 2nd edn. Cambridge, MA: The MIT Press; 2018.
37. Bartolucci R, Grandoni S, Melillo N, et al.Artificial intelligence and machine learning: just a hype or a new opportunity for pharmacometrics? In Page Meet. 28 Stock. Abstract 9148. 2019.
38. Ribba B, Dudal S, Lavé T, Peck RW. Model-informed artificial intelligence: reinforcement learning for precision dosing. *Clin Pharmacol Ther*. 2020;107:853–857.
39. Henrich A. Pharmacometric modelling and simulation to optimise paclitaxel combination therapy based on pharmacokinetics, cumulative neutropenia and efficacy. PhD thesis, Freie Universität Berlin. 2017. https://doi.org.https://doi.org/10.17169/refubium-12511.
40. Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem. *Mach Learn*. 2002;47:235–256.
41. Coulom R. Efficient selectivity and backup operators in Monte-Carlo tree search. In *5th Int. Conf. Comput. Games*, 72—83. 2007.
42. Kocsis L, Szepesvári C. Bandit based Monte-Carlo planning. *Machine Learning: ECML*. 2006;4212:282–293. http://link.springer.com/10.1007/11871842_29
43. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nat Publ Gr*. 2017;550:354–359.
44. Sutton RS. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bull*. 1991;2(4):160–163.
45. Silver D, Sutton RS, Martin M. Sample-based learning and search with permanent and transient memories. In *Proc. 25th Int. Conf. Mach. Learn. Helsinki, Finl*. (2008). https://doi.org/10.1145/1390156.1390278.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: 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;10:241–254. https://doi.org/10.1002/psp4.12588