UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXXVII CICLO - 2024

REINFORCEMENT LEARNING AND PK-PD MODELLING FOR PRECISION DOSING

PhD Thesis by

ALESSANDRO DE CARLO

Advisors:
Prof. Paolo Magni
Prof. Elena Tosca

PhD Program Chair:
Prof. Silvana Quaglini

Abstract (Italiano)

Il *Precision Dosing* è approccio clinico all'interno del panorama della medicina di precisione, in grado di superare le limitazioni del classico paradigma clinico del *one-size-fits-all*. Questo metodo paziente-centrico ha attirato l'interesse sia dei clinici che delle agenzie regolatorie per la sua capacità di adattare una terapia farmacologica al singolo soggetto, ottimizzandone gli *outcomes*. La sua applicazione è di particolare rilevanza in contesti particolarmente complessi caratterizzati dalla somministrazione di farmaci con una finestra terapeutica ristretta, da una significativa variabilità interindividuale (IIV), dalla presenza di gravi effetti avversi (AEs) derivanti da sovradosaggio o dal coinvolgimento di popolazioni speciali. In tali contesti clinici, una delle principali tecniche adottate dal precision dosing sono le strategie di dosaggio adattativo che sfruttano il monitoraggio di biomarcatori di efficacia e tossicità per personalizzare la terapia farmacologica.

I modelli di farmacometria assumono un ruolo fondamentale nel supportare i processi decisionali tipici del precision dosing, e l'approccio *Model Informed Precision Dosing* (MIPD) è stato proposto per diversi composti. Il recente interesse verso l'intelligenza artificiale e il machine learning (AI/ML) nel campo della farmacometria offre l'opportunità di esplorare nuove metodologie ibride che integrano modelli PK-PD con tecniche di AI/ML per migliorare il MIPD in presenza di dosaggi adattativi. Tra queste, il Reinforcement Learning (RL), una sotto branca del ML caratterizzata da algoritmi per risolvere problemi decisionali sequenziali, è attualmente di particolare interesse. Infatti, il RL si adatta perfettamente ai problemi di dosaggio di precisione e alle strategie di dosaggio adattativo che coinvolgono il monitoraggio periodico del paziente. Di conseguenza, l'integrazione dei modelli PK-PD con tecniche di RL, denominata *Model-Informed RL* (MIRL), è attualmente uno degli argomenti di principale interesse in farmacometria.

L'obiettivo di questa tesi è di esplorare le potenzialità di tecniche ibride basate sulla combinazione di RL e modelli PK-PD per supportare i problemi di dosaggio di precisione, ponendo possibili soluzioni per le limitazioni degli approcci esistenti nella letteratura. In particolare, è stato sviluppato un nuovo framework di MIRL per apprendere strategie di dosaggio adattativo clinicamente accettabili e personalizzate per ciascun paziente. Tale approccio è stato applicato e valutato su diversi problemi di dosaggio di precisione tratti da casi di studio reali, comprendenti sia farmaci approvati sia in fase di sviluppo. In questa tesi è stato inoltre proposto un riadattamento dell'approccio MIRL di letteratura per ricavare strategie di dosaggio generali

per un'intera popolazione di pazienti. In questo modo, è stato possibile comparare in maniera più robusta le performances della nuova metodologia.

Dopo una breve introduzione ai concetti fondamenti relativi al precision dosing (Capitolo 1), nel Capitolo 2 viene fornita la formalizzazione del workflow di MIRL sviluppato per ottenere protocolli di dosaggio adattivi specifici per ogni paziente. Nel corso della trattazione, sono state evidenziate anche le differenze di tale workflow rispetto agli approcci RL/PK-PD già esistenti in letteratura. Il nuovo framework MIRL usa il Q-learning (QL) come algoritmo di RL, introducendo così un agente QL personalizzato (QLind) addestrato sullo specifico paziente. Questo approccio ibrido RL/PK-PD sfrutta un *digital twin* del paziente per fornire l'esperienza necessaria ad addestrare gli agenti QLind a ricavare strategie di dosaggio personalizzati. Il *digital twin* corrisponde a un modello PK-PD del paziente caratterizzato da un insieme di parametri individuali e covariate. Questo framework consente un livello di personalizzazione più alto rispetto agli approcci MIRL tradizionali, che in genere si basano sull'addestramento di un singolo agente di RL su un'intera popolazione di pazienti, ottenendo così protocolli di dosaggio adattativo generali.

L'approccio di MIRL incentrato sul singolo paziente è stato valutato su tre problemi reali di precision dosing aventi complessità crescente, (capitoli 3-5). Tali valutazioni sono state caratterizzate da semplificazioni necessarie per ottenere una valutazione più chiara della metodologia. È stato infatti ipotizzato che i modelli PK-PD dei pazienti descrivessero perfettamente la farmacoterapia, andando dunque a trascurare la variabilità residua non spiegata (RUV) dovuta a inesattezze del processo di modellazione. È stato inoltre supposto di conoscere esattamente tutti i parametri dei modelli PK-PD che descrivono i singoli pazienti prima dell'inizio del trattamento.

Sotto queste ipotesi, i protocolli di dosaggio adattativo ricavati dal MIRL per ogni soggetto hanno raggiunto ottime prestazioni in tutti e tre i casi di studio. Queste strategie di dosaggio basate sul RL e personalizzate sul singolo soggetto, sono state in grado di ottimizzare su ciascun paziente diversi outcomes clinici, inclusi quelli a lungo termine (ad esempio, la probabilità di sopravvivenza del paziente), sia per monoterapie sia per somministrazioni concomitanti di più farmaci.

Infine, nel Capitolo 6 sono state proposte due estensioni dell'approccio di MIRL introdotto in questa tesi per superare le ipotesi semplificative utilizzate nelle valutazioni sopracitate che, attualmente, rappresentano un limite all'applicazione delle tecniche di RL/PL-PD nella pratica clinica. In primo luogo, il framework di MIRL è stato calato in un contesto bayesiano per superare l'assunzione che il digital twin del paziente debba esser completamente noto prima dell'inizio del trattamento. In secondo luogo, è stata presentata una variante del QL, denominata EQL, che sfrutta le simulazioni Monte Carlo di modelli PK-PD per considerare esplicitamente la RUV dovuta alle approssimazioni introdotte dal modello.

I risultati ottenuti hanno dimostrato che questi nuovi rifinimenti del paradigma MIRL incentrato sullo specifico paziente possono superare con successo le principali limitazioni di questo workflow, favorendo dunque una sua applicazione nella pratica clinica attuale.

Nonostante siano necessarie ulteriori investigazioni, i risultati presentati in questa tesi evidenziano le potenzialità degli approcci MIRL per supportare una vasta gamma di scenari di precision dosing e propongono alcune soluzioni interessanti per accelerare la loro integrazione a supporto della farmacologia clinica di ogni giorno.

Abstract (English)

Precision dosing is a revolutionary clinical approach within the precision medicine landscape which allows to overcome the current limitations of the classical one-size-fits-all approach. This patient-centric approach earnt lot of attention from both the clinical community and regulatory agencies due to its potentialities to optimize therapeutic outcomes for drugs with a narrow therapeutic window, significant inter-individual variability (IIV), severe adverse effect (AEs) from overdosing or administered in special populations. One of the main frameworks within precision dosing landscape are adaptive dosing strategies which leverage the monitoring of efficacy and toxicity biomarkers specific to each patient in order to customize the pharmacotherapy.

Pharmacometrics modeling is central to support precision dosing problem, and the Model Informed Precision Dosing (MIPD) approach has rapidly gained momentum. The recent outbreak of Artificial Intelligence and Machine Learning (AI/ML) within pharmacometrics presents an opportunity to explore novel hybrid methodologies integrating PK-PD modelling with AI/ML techniques to improve MIPD in adaptive dosing context. Among these, Reinforcement Learning (RL), a ML subfield characterized by algorithms for solving sequentially precision dosing problems, gained lot of interest. Indeed, RL naturally fits to precision dosing problems involving periodic patient monitoring and adaptive dosing strategies. Consequently, the integration of PK-PD modeling with RL, referred to as Model-Informed RL (MIRL), is currently at the core of the debate in pharmacometrics to support precision dosing.

The aim of this thesis is to explore the potentialities of the RL/PK-PD framework to support precision dosing problems, addressing the limitations of existing approaches in literature. In particular, a novel MIRL framework was here developed to learn clinically acceptable adaptive dosing strategies tailored to individual patients. It was applied and evaluated on several real-world-based precision dosing problems, including both approved and under-development drugs. Furthermore, a literature MIRL framework was adapted in this dissertation to automatically derive general, clinically acceptable adaptive dosing rules for patient populations, allowing for a more comprehensive understanding of the potential and challenges of MIRL approaches.

A brief introduction of main precision dosing concepts is provided in Chapter 1. Then, Chapter 2 introduces a formal description of the patient-specific MIRL workflow developed in this thesis, highlighting also its differences from existing RL/PK-PD approaches. The novel MIRL

framework employs Q-learning (QL) as RL algorithm, thus introducing a personalized QL-agent (QLind) trained for each specific patient. This hybrid RL/PK-PD approach leverages patient digital twins to provide the experience necessary to train QLind-agents for solving precision dosing tasks. These individual virtual replicas correspond to patient PK-PD models characterized by a set of individual model parameters and covariates. This framework allows a deeper level of personalization than traditional MIRL approaches, which typically rely on training a single RL-agent on an entire patient population, resulting in general adaptive dosing protocols.

The patient-centric MIRL framework was evaluated on three real precision dosing problems of increasing complexity, presented in Chapters 3-5. These evaluations were characterized by necessary simplifications for a clearer assessment of the methodology. It was hypothesized that patient PK-PD models perfectly describe the pharmacotherapy and, residual unexplained variability (RUV) due to model misspecification was neglected. Furthermore, a full knowledge of individual patient PK-PD parameters prior to treatment start was assumed.

Under these assumptions, individually tailored MIRL adaptive dosing protocols achieved very good performances on the three case studies. Such RL-based strategies were able to optimize on each patient various treatment outcomes, including long-term ones (e.g., patient survival probability), for both monotherapies and concomitant drug administrations.

Finally, in Chapter 6 two extensions of the MIRL approach were developed to overcome the assumptions used in this evaluation that represent a limit to a practical application of MIRL within clinical setting. First, a Bayesian MIRL framework was developed to account for the impractical assumption of fully characterized patient digital twins before treatment begins. Secondly, a variant of QL algorithm combined with PK-PD Monte Carlo simulations, (EQL), was presented to explicitly tackle RUV due to model misspecifications. The obtained results demonstrated that these techniques can successfully address key limitations of the individual-oriented MIRL paradigm.

Although further research is needed, the results presented in this thesis highlight the potentialities of MIRL approaches to support a wide range of precision dosing tasks and propose some interesting solutions to move them closer to real-world clinical applications.

Contents

Introduction **9**
1.1. Precision dosing workflow *10*
1.2. Model-Informed Precision Dosing (MIPD) *11*
1.3. Reinforcement Learning for precision dosing *14*
1.4. Thesis Overview *15*

Integrating RL and PK-PD models in the precision dosing context **17**
2.1. Reinforcement Learning *17*
2.1.1. Markov Decision Processes 18
2.1.2. Policy and Value Functions 20
2.1.3. Q-Learning algorithm 21
2.2. Tackling a precision dosing problem with RL and PK-PD models *23*
2.2.1. Learning an optimal adaptive dosing protocol for an entire patient population with RL and PK-PD models 27
2.2.2. Learning patient-specific adaptive dosing strategies with RL and PK-PD models 29

Optimizing a single efficacy/toxicity endpoint. Application of the RL/PK-PD framework to the erdafitinib precision dosing problem **32**
3.1. Methods *33*
3.1.1. Erdafitinib case-study 33
3.1.2. RL-based precision dosing of erdafitinib 34
3.1.2.1. Reward function 35
3.1.2.2. System/Patient health state 36
3.1.2.3. Agent Actions 38
3.1.2.4. Implementation 39
3.1.2.5. Evaluation setup 40
3.2. Results *40*
3.2.1. RL-based personalized dosing strategies vs FDA-approved protocol 41
3.2.2. RL general protocol vs individual RL-agents 51
3.3. Discussions *55*

Joint optimization of multiple treatment biomarkers. Application of the RL/PK-PD framework to givinostat therapy in polycythemia vera patients **59**
4.1. Methods *60*
4.1.1. Givinostat treatment of polycythemia vera 60
4.1.2. Setup of QL algorithm for givinostat precision dosing 62
4.1.2.1. Reward function 63
4.1.2.2. System/Patient states 65
4.1.2.3. QL-Agent actions 66
4.1.2.4. Implementation 68
4.1.2.5. Learning a unique adaptive dosing protocol for the whole population with QL 69
4.1.2.6. Learning patient-specific adaptive dosing strategies with QL 69
4.2. Results *70*
4.2.1. Learning a unique adaptive protocol for the whole population with QL 70

4.2.2. Learning patient-specific adaptive dosing strategies with QL 74
4.2.3. Adapting patient-specific QL-based protocols to optimize givinostat phase III trial 80
4.3. Discussions *82*

Optimization of short- and long-term outcomes of co-administered drugs. Application of the RL/PK-PD framework to axitinib/anti-hypertensive treatment in advanced renal cancer 85
5.1. Methods *86*
5.1.1. Axitinib /anti-hypertensive treatment in advanced renal cancer 86
5.1.2. Set up of QL algorithm for axitinib/anti-hypertensive co-administration 88
5.1.2.1. Short-term reward function 89
5.1.2.2. Short- and long-term reward function 91
5.1.2.3. System/patient states 92
5.1.2.4. QL Agent actions 93
5.1.2.5. Evaluation framework 94
5.2. Results *95*
5.2.1. Treatment personalization based on short-term reward function 95
5.2.2. Integrating long-term outcomes in the reward function 99
5.3. Discussions *108*

Overcoming Key Challenges in RL/PK-PD Framework for Clinical Integration 113
6.1. Integration of RL/PK-PD framework with Bayesian estimation 114
6.1.1. Application of the Bayesian RL/PK-PD to givinostat case-stdy 115
6.1.2. Results 118
6.2. Extending the RL/PK-PD framework to a stochastic treatment response. 122
6.2.1. EQL algorithm 125
6.2.2. Continuous vancomycin infusion regime in ICU patients 126
6.2.3. Formalization of vancomycin precision dosing problem within MDP framework 127
6.2.3.1. Reward function 128
6.2.3.2. System/Patient states 129
6.2.3.3. EQL Agent actions 130
6.2.3.4. Evaluation Framework of EQL agent 131
6.2.3.5. Results 132
6.3. Discussions *137*

Overall Conclusions 142

Overview of Reinforcement Learning Algorithms 147
A.1. Monte-Carlo Tree Search *147*
A.2. Deep Q-Learning *150*
A.3. Actor critic DQL *152*

Supplementary Materials of Chapter 3 155
B.1. Erdafitinib PK-PD modelling *155*
B.1.1. Population PK model of Erdafitinib 155
B.1.2. Population PK-PD model of erdafitinib 159
B.1.3. Model analysis 162
B.1.4. Mlxtran code for erdafitinib PK-PD model 163
B.2. Virtual patient population to evaluate the RL/PK-PD approach *165*
B.2.1. Statistical distributions of covariates 165
B.2.2. Generation of the preliminary virtual population 165
B.2.3. Inclusion criteria for completely responsive patients 166
B.2.4. Inclusion criteria for partially responsive patients 169

B.3. Hyperparameters of QL algorithm *171*

Supplementary Materials of Chapter 4 ..... **172**

C.1. Terms in the reward function evaluating the derivative of the haematological parameters *172*

C.2. Givinostat PK-PD modelling framework *174*

C.2.1. Population model of givinostat PK 174

C.2.2. PK-PD modelling of givinostat effect on PLT, WBC and HCT 175

C.2.3. Steady-state analysis of givinostat PK-PD model 177

C.3. Generating a virtual population of Polycythemia Vera patients *178*

C.4. QL algorithm hyperparameters *181*

C.5. Tuning the reward function to learn a unique adaptive dosing protocol for the whole population with QL *182*

Supplementary Materials of Chapter 5 ..... **185**

D.1. Empirical Pop-PK-PD-OS model for the co-administration of axitinib and anti-hypertensives *185*

D.1.1. PK model of axitinib daily exposure 186

D.1.2. PK-PD model of axtinib effect of dBP 187

D.1.3. Empirical AX-AH PK-PD model of dBP 188

D.1.4. PK-PD model of axitinib effect on sVEGFR-3 188

D.1.5. PK-PD model of axitinib effect on SLD 189

D.1.6. PK-PD-OS model of axitinib 190

D.1.7. Assumption on AH medications following steady-state analysis of dBP PK-PD model 190

D.1.8. Steady-state analysis of SLD PK-PD model 191

D.2. Generation of the virtual test population *192*

D.2.1. Statistical distributions of patient covariates 192

D.2.2. Stratified random sampling to generate the virtual patient population 193

D.3. Hyperparameters of QL algorithm *195*

D.4. Supplementary figures for QLind-agents trained with S&LT-Reward function *196*

Supplementary Materials of Chapter 6 ..... **198**

E.1. Simplified version of givinostat precision dosing problem *198*

E.2. Generation of the virtual population *199*

E.3. QLpop-agent used in the Bayesian MIRL approach *200*

E.4. Hyperparameters of QLind and QLind-bay agents *202*

E.5. Prioritized Sweeping QL *202*

E.5. Vancomycin Pop-PK model *203*

E.6. Algorithm Hyperparameters *204*

References **205**

Chapter 1

Introduction

The concept of *personalized medicine* dates back to the Greek era. Already Hippocrates remarked in his writings on the importance of customizing pharmacological treatment according to individual characteristics [1,2]. Twenty-five centuries after the father of modern medicine, the breakthrough of human genome mapping [3] generated new interest in personalized therapies leveraging the knowledge of patient genetic makeup [4–6]. Starting from the 2010s, *personalized medicine'* was extended by the broader notion of '*precision medicine*'. As defined by the US National Research Council, precision medicine refers to an emerging healthcare approach tailoring disease prevention and treatment on patient characteristics [7–9]. In this framework, multimodal information including not only genomics but also laboratory results and imaging are leveraged to identify the most effective treatment plan for each individual [7,9].

Precision medicine was definitively brought into the spotlight when US president Barack Obama launched the $215 million '*Precision Medicine Initiative'* in 2015 [9,10]. Following this announcement, the interest in precision medicine grew exponentially [11–13], reaching more than 30,000 publications annually in the PubMed database from 2020 onward [14]. In particular, the use of Artificial Intelligence (AI) and Machine Learning (ML) in supporting precision medicine was largely investigated in these works [15–18]. Indeed, AI and ML algorithms are well-suited to the precision medicine paradigm due to their ability to process large amounts of data, automatically detect patterns, and use these uncovered patterns to make inferences [19]. Several works highlighted promising results of AI/ML models improving diagnosis, risk prediction and patient stratification, thereby fostering precision medicine in delivering personalized healthcare assistance [15,17,18].

Precision medicine is not limited to provide detailed early-diagnosis and to identify the most appropriate drug for a specific patient having a certain disease. In fact, once the best compound for that patient has been identified, the focus of this patient-centric approach shifts to define a tailored administration protocol maximizing the efficacy-safety trade-off [20]. This stage pertains to a branch of precision medicine known as *'precision dosing'*

[21–26]. In this field, pharmacometrics mathematical modelling and simulations (M&S) characterizing drug pharmacokinetics and pharmacodynamics (PK-PD) phenomena are the core of clinical decision-making support [27,28]. Their integration with AI/ML algorithms is currently under investigation to further optimize dose personalization for several classes of molecules [29,30].

This thesis deals with the development of a hybrid framework combining PK-PD M&S with a class of AI/ML algorithms, Reinforcement Learning (RL), to support the precision dosing process. Several applications to real precision dosing problems having different degrees of complexity will be presented in this work. Furthermore, current limitations and possible solutions towards a clinical application will be discussed.

1.1. Precision dosing workflow

Precision dosing refers to tailoring the administration of a drug to a specific patient at a given time considering individual factors known to alter drug disposition and/or response [22,26,31]. The aim of this approach is to maximize treatment effectiveness in each patient and, simultaneously, its safety by reducing the onset of drug-related adverse events [20,21,32].

Accounting for individual characteristics in the formulation of a dosing regimen is crucial for some compounds because different patients may respond in an extremely different way to the same dose (i.e., inter-individual variability) and also the same patient may respond differently to the same dose over time (intra-individual variability). In such cases, the traditional *'one-size-fits-all'* paradigm, which follows a dosing protocol that works well for the majority of patients, can lead to incorrect dosing, thus provoking serious consequences, including lack of efficacy, uncontrolled disease, and intolerable, potentially life-threatening adverse effects [22,26].

Consequently, the interest in precision dosing has significantly increased both within the clinical community and among regulatory agencies (e.g., US Food and Drug Administration, FDA) and industries [24,26–28,31,33]. Indeed, such approach can potentially improve also the drug development process by reducing the attrition rates for those molecules showing inadequate efficacy/toxicity performances with the *'one-size-fits-all'* approach [26,32].

Therefore, precision dosing has been recommended for those compounds showing a narrow therapeutic window, significant inter/intra-individual variability, severe or irreversible adverse effects due to overdosing and/or for diseases with serious consequences of undertreatment. Moreover, guidelines proposed this approach in presence of special populations (e.g., elderly, critically ill, children) or invasive routes of administration [22,24,25,32,34–36]. Actually, it is reported in the literature that precision dosing is required for different classes of compounds such as antimicrobial,

anticancer, biologics, antiretrovirals, psychotropic, immunosuppressants, anti-coagulants and anticonvulsants [21,26,31,37].

Adaptive-dosing strategies are a particular precision dosing workflow in which dose need to be repeatedly adjusted to a time-varying patient condition [38,39] (Figure 1). In these cases, patients are periodically monitored to assess disease status and to measure drug plasma concentration (i.e., therapeutic drug monitoring, TDM), clinical response or, in most of the cases, levels of some efficacy and/or safety biomarkers [34,37,40–43]. These information as well as patient clinical history, are used by clinicians to modify drug dose by leveraging an available dosing algorithm (e.g., reported in clinical practice guidelines [44]). Therefore, an adaptive dosing regimen is formalized by a set of decision rules that specify how the dose (or, more generally, the treatment) should be repeatedly adjusted through time in response to an evolving patient profile [45,46].

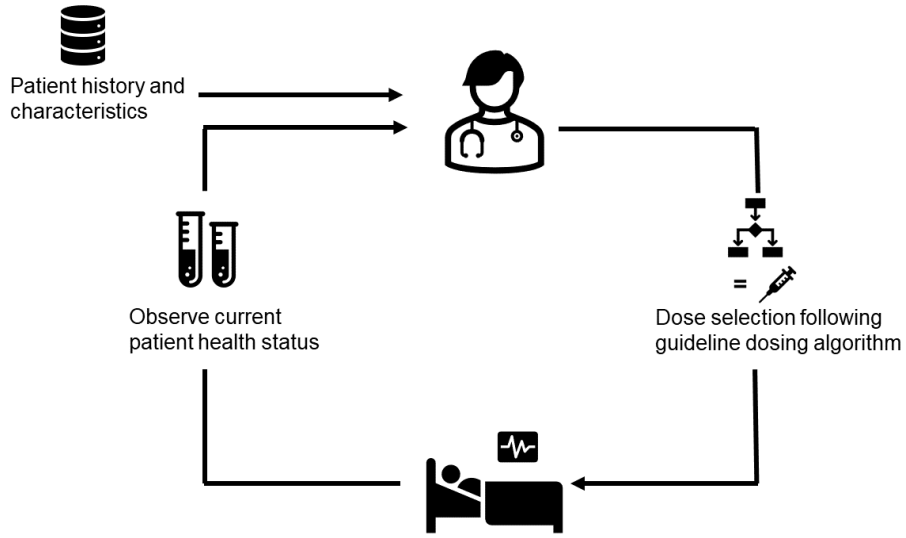


Figure 1: The workflow of precision dosing based on adaptive dosing strategies. Patient health status is periodically monitored by performing measurements of treatment efficacy/toxicity biomarkers as well as assessments of disease degree. Then, clinicians integrate this information with previous patient history (e.g., collected into electronic health records) to select the dose for the next treatment cycle. This decision is taken considering pre-defined dosing algorithms (e.g., available from guidelines).

1.2. Model-Informed Precision Dosing (MIPD)

The idea that M&S of drug PK-PD could inform personalized adaptive dosing strategies dates back to the late 1960s with the works of Sheiner and Jelliffe [27,47–49]. Fifty years later, with the rise of precision medicine and the rapid expansion of pharmacometrics, this approach gained great momentum becoming a “state-of-art” method under the name of model-

informed precision dosing (MIPD)[27,28,49]. Indeed, several commercial software tools are nowadays available to enable MIPD for different drug classes [27,50] and different successful implementations of this framework are reported in the literature [51–54].

The core of MIPD is represented by population PK-PD models (Pop-PK-PD) which quantitatively describe the PK-PD dynamics of a drug within a population of patients [27,55,56]. To this end, Pop-PK-PD models are built leveraging the widespread and well-established framework of non-linear mixed effect models (NLMEM) [27,56].

In particular, NLMEM are composed by a structural (i.e., deterministic) component which semi-mechanistically describes PK-PD processes with parameters-dependent ordinary differential equations. Some stochastic components are then added to describe the random aspects of the PK-PD process such as inter/intra-individual variability and residual unexplained variability (RUV) [56,57]. Inter-individual variability (IIV) describes the random differences between subject of the same population. Part of the IIV is typically explained in the NLMEM framework by a covariate model describing the impact of patients' characteristics on PK-PD mechanisms. A further layer of randomness is intra-individual variability or inter-occasional variability (IOV) that accounts for differences in the PK-PD response occurring between different treatment episodes. Differently, RUV represents the unexplained shift between model predictions and observations, which can be imputable to noise and/or model misspecifications [56–59].

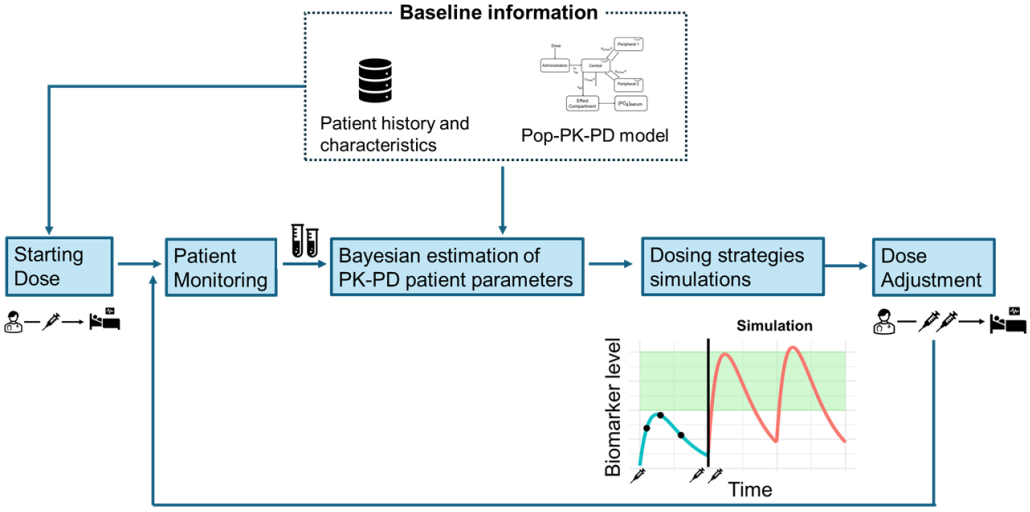


Figure 2: A schematical representation of the MIPD paradigm supporting adaptive dosing strategies. Patient history and characteristics (e.g., covariates) and Pop-PK-PD models already developed on large populations are used as prior information which can inform the selection of the starting dose. Then, efficacy/toxicity biomarkers are monitored during treatment and observations are merged with prior information to update patient PK-PD model parameters with Bayesian estimation. Once the model is updated, it can be leveraged to perform simulations of different dosing scenarios for the next treatment cycles. Thus, the most likely next-cycle dosage maintaining the biomarkers in the target range is identified and used to inform dose adjustment.

In the MIPD paradigm, Pop-PK-PD models are generally combined with a Bayesian framework which allows the estimation of patient model parameters by integrating observed data with prior knowledge. As illustrated in Figure 2, Pop-PK-PD models already developed on large datasets are used as *prior* knowledge which can be used to inform the selection of the initial dose. During pharmacotherapy, monitored individual characteristics (e.g., drug concentration, biomarker levels) are merged with the prior information to obtain a Bayesian estimation of patient PK-PD model parameters and their relative uncertainty. Patient PK-PD model and its parameters set constitute a *digital twin* of the real patient [60,61]. Consequently, once this virtual replica of the patient is updated with the Bayesian estimation, simulations of different dosing scenarios are performed to identify the most appropriate dosing schedule to reach the targeted endpoints [27,49,62–64].

Another possible usage of Pop-PK-PD models within the MIPD context is to define and/or fine-tune a general adaptive dosing protocols for specific populations or sub-populations of interest (e.g., stratified by patients covariates). In this case, a trial-and-error-procedure is performed and Pop-PK-PD model simulations are used to evaluate both the efficacy and safety of the candidate adaptive dosing algorithm [27,45,65,66]. Therefore,

simulation results are used to inform changes in the dosing protocols or comparing different strategies.

1.3. Reinforcement Learning for precision dosing

The recent outbreak of AI/ML in pharmacometrics represents an opportunity to explore novel methodologies to further improve MIPD in the context of adaptive dosing strategies[27,67–70]. In particular, RL is a branch of ML which naturally fits to precision dosing problem based on periodic patient monitoring and adaptive dosing strategy [39,71–75].

Indeed, RL includes a set of algorithms for solving sequential decision-making problems, where an agent (i.e., the decision-maker) interacts with a system over time to steer it toward a given target condition. RL algorithms aim to estimate an optimal policy that allows the agent to select, at each system state, the best action to achieve the desired target condition on the system [76]. The optimal policy is learnt through a trial-and-error interplay between system and agent: at each system state, the agent tries different actions on the system and evaluates their consequences (i.e., new system state) based on a reward signal quantifying the suitability of the performed action to achieve the final goal [76]. In the context of precision dosing, the agent represents the clinician that periodically monitors the patient status through efficacy/safety biomarkers (i.e., system state) and consequently selects the dose to administer (i.e., action). Patient response in terms of biomarker level (i.e., new system state) is observed and the suitability of the administered dose (i.e., selected action) is quantified based on distance from the target (i.e., collection of the reward) [30,39].

Therefore, when applied to a precision dosing problem in which the dose needs periodic adjustments according to patient monitoring, RL algorithms have the potential to learn the optimal dosing strategy to achieve the targeted therapeutic goal. Consequently, the interest around this ML subfield has rapidly grown among pharmacometricians and its potentialities and pitfalls are currently at the core of the debate [39,71,74].

RL algorithms require a trial-and-error procedure based on trying different actions/doses at each system/patient state to estimate the optimal dosing policy. Due to obvious ethical concerns, RL algorithms for precision dosing cannot be directly developed on actual patients as this would mean exposing a large number of patients to potentially suboptimal or, even, dangerous treatments [39].

A first approach adopted in different literature works was based on training RL algorithm on retrospective clinical data from previously treated patients [30,39,71,77,78]. However, the amount of data needed to adequately train a RL algorithm is extremely high. For example, in [77] 170,000 clinical records were used to estimate the optimal dosing policy of antiseptic medications in intensive care unit patients. In addition, learning datasets should include a multitude of dosing scenarios, even those particularly

harmful for patient safety (e.g., extreme over/under dosing conditions). Datasets with these characteristics are cannot generally be collected and made available both in the clinical and the clinical trial settings [39]

To overcome these issues, different works have underlined the importance of integrating Pop-PK-PD models within the RL framework, thus funding a novel approach i.e., Model-Informed Reinforcement Learning (MIRL) for precision dosing [30,39,73,79,80]. Model simulations are central in this hybrid technique as they compensate for the lack of real-world experience, thus allowing a better training of RL-agents based on several simulated dosing scenarios, including those life-threatening [73,79].

Preliminary literature works used MIRL to derive an optimal dosing strategy for an entire patient population which translates into having a single RL algorithm trained to optimize treatment for all the patients in the population [39]. More in details, such approach was applied to optimize propofol-induced anesthesia [73,81–84], anticancer treatments [74,75,85–88], anticoagulant therapy with warfarin [89,90], insulin injection in diabetics [91] and hemodialysis in anemia patients [92].

The obtained results shown that protocols designed by MIRL reached equivalent or superior performances with respect to guidelines dosing strategies (i.e., formulated by expert clinicians) and traditional control algorithms (e.g., proportional-integral-derivative controller, PID) [83]. Interestingly, it was demonstrated that MIRL works in average well in populations characterized by low to moderate IIV [39]. However it was demonstrated that in case of high IIV, the RL-based precision dosing strategies could work sub-optimally or inefficiently [39,73]. Some additional limitations affect the current MIRL approach. For example, available studies on MIRL are often limited by a simplified representation of the precision dosing problem leading to infeasible dose suggestion due to clinical unacceptability [39,79]. Furthermore, in most of these works, RL was challenged to optimize only one treatment efficacy end/or toxicity endpoint, neglecting other potentially relevant biomarkers of clinical interest [74]. Some works used a simplified structural PK-PD model while in others not all sources of random variability (i.e., RUV, IIV, and when present, IOV) where considered [39,74]. Finally, there is still lack of knowledge on the performances of MIRL on the optimization of both long-term outcomes (e.g., overall survival) and concomitant drugs administration [74].

1.4. Thesis Overview

Following the above considerations, the aim of this thesis is to explore the integration of RL with PK-PD models to deliver precision dosing, addressing the limitations of the works currently available in the literature. In this work, a novel MIRL framework to learn clinically acceptable adaptive dosing strategies tailored on each patient will be presented and evaluated on real precision dosing problems throughout the work. Both already approved and under development drugs will be considered to show also the potentialities

of this method in the clinical trials context. In this investigation, the literature MIRL framework [73,81–85,88,89,92] will be adapted to derive a set of general clinically acceptable adaptive dosing rules (i.e., a single RL-based controller for all the patient population) on some of the proposed precision dosing scenarios to better understand its potentialities and challenges. Furthermore, this strategy will be used as benchmark for the novel MIRL framework to provide a more robust assessment of its potentialities.

In Chapter 2, the focus will be on presenting the methodologies implemented in this thesis. First, the mathematical background of RL and the main algorithm used in thesis, Q-Learning, will be discussed. Then, the attention will be moved to the novel implemented MIRL methodology for precision dosing and its differences with the literature-based approach. Appendix A extends this chapter by providing an overview of the other RL algorithms available.

Chapters 3 to 5 and their respective appendices (B-D) provide a detailed evaluation of the implemented MIRL workflow on three real precision dosing problems of increasing complexity.

First, in Chapter 3, erdafitinib precision dosing will be leveraged to demonstrate how this novel framework can be used to optimize a single efficacy and toxicity endpoint with personalized dosing strategies. Then, in Chapter 4, the focus will shift to challenging the MIRL methodology to jointly optimize the multiple safety/efficacy biomarkers of givinostat precision dosing in polycythemia vera patients. Chapter 5 will present an application of the proposed MIRL approach to optimize short and long-term outcomes in presence of concomitant drugs administration. To this end, the axitinib/anti-hypertensive precision dosing problem will be considered as case study.

In Chapter 6 main aspects limiting the application of the proposed method within the clinical practice and possible solutions to circumvent them will be discussed through different applications. Information supporting this chapter will be in Appendix E.

Finally, an overall conclusion is reported in Chapter 7.

Chapter 2

Integrating RL and PK-PD models in the precision dosing context[1]

The aim of this chapter is presenting how RL and PK-PD models can be coupled to optimize adaptive dosing protocols. To this end, first, the mathematical background of RL will be described in sections 2.1.1. and 2.1.2. following the dissertation provided by A.S. Sutton and A.G. Barto in [76]. Secondly, the RL algorithm used in this thesis, Q-Learning (QL), its numerical implementation and technical advantages will be the core of section 2.1.3. Further details on other RL algorithms available in literature are reported in Appendix A.

After this theoretical background on RL and QL, the MIRL methodologies implemented in this thesis will be extensively covered in section 2.2. In particular, section 2.2.1. will illustrate an adaptation of the most common literature MIRL approach to derive general adaptive dosing strategies within a target population. Finally, in section 2.2.2., the focus will be on the developed novel MIRL paradigm which provides a deeper level of treatment personalization by learning a patient-specific adaptive dosing protocol.

2.1. Reinforcement Learning

RL is a ML branch including a set of algorithms to solve sequential decision-making processes in which a decision-maker (i.e., the agent) aims to control a system. Therefore, the goal of RL methods is to learn an optimal decisional policy that allows the agent to select, at each system state, the best action to drive the system towards a target condition. To this end, RL emulates learning from interaction process performed by humans and other animals [72,76]. Indeed, RL algorithms use a trial-and-error procedure in

[1] The content of this chapter was published in [39].

which the agent, at each system state, tries different actions on the system. Then, a reward is returned to the agent by the system to evaluate the consequences of agent's actions (i.e., new system state). This feedback signal is crucial to learn the optimal policy as it teaches which action, depending on the current system state, is the most appropriate to achieve the final goal.

This qualitative description of RL algorithms will be mathematically formalized in sections 2.1.1. and 2.1.2.

2.1.1. Markov Decision Processes

The mathematical framework of RL is represented by a finite Markov Decision Process (MDP) which describes a sequential decision-making scenario in a discrete domain [76,93,94]. This original formulation was also extended to a continuous space as detailed in [95]. Figure 3 provides a graphical representation of a finite MDP, in which the agent, representing the decision-maker, interacts with the system (alternatively named environment) to steer it towards a target condition. In particular, it is assumed that the agent-system interplay occurs at each step of a finite time sequence, $t = 0,1,2,\cdots,T$. At each step t, the agent observes the system state, $S_t = s$, and, consequently, performs an action, $A_t = a$, on the system. At step $t+1$, the system evolves to the next state, S_{t+1}, and the agent receives a reward, R_{t+1}, in response to its previous action. The reward signal quantifies the goodness of the agent action with respect to the final aim. Therefore, a higher R_{t+1} is returned by system whether performing $A_t = a$ when $S_t = s$ drove the system closer to the target condition.

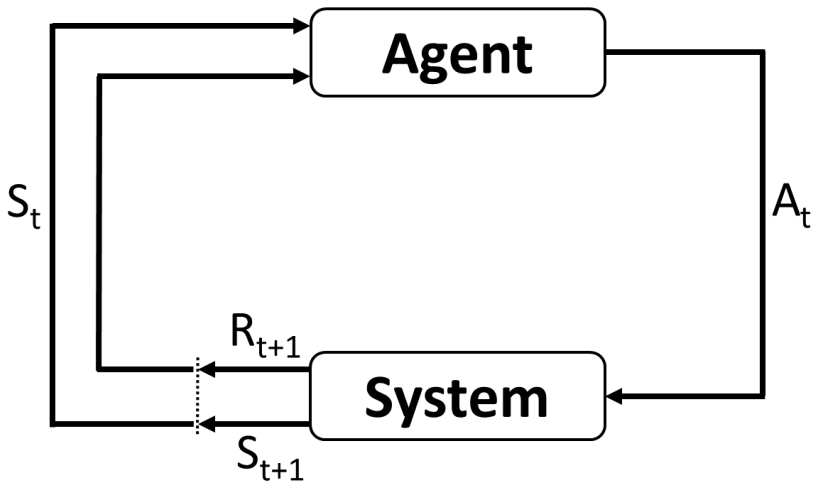


Figure 3: Schematical representation of a Markov Decision Process (MDP).

A finite MDP can be mathematically described by the following elements:

- $S = \{s_1, \cdots, s_N\}$, a finite set of N states describing the system;
- $A = \{a_1, \cdots, a_M\}$, a finite set of M actions that the agent can perform on the system;
- $T: S'\ x\ A\ x\ S \rightarrow [0,1]$, a Markovian state transition probability matrix where, $T(s', a, s) = P\{S_{t+1} = s' | S_t = s, A_t = a\}$ is the probability that the system evolves to state s' at time $t + 1$, given the current state s and action a;
- $\rho: S'\ x\ A\ x\ S \rightarrow \mathbb{R}$ is the reward function where $\rho(s', a, s) = R_{t+1}$ is the reward that the system returns at time $t + 1$ to the agent by evolving from the state $S_t = s$ to $S_{t+1} = s'$ because of action $A_t = a$.

When repeated for each $t = 0,1,2, \cdots, T$, the sequential agent-system interaction of Figure 3 generates a sequence or *trajectory* of states, actions and rewards:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots.$$

(1)

As the Markov property holds, the probability that at time step $t + 1$ the system evolves to state s' and returns the reward r only depends on the system state and action at time t. Therefore, the dynamics of a MDP are fully specified by:

$$p(s', r | s, a) = P\{S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\}.$$

(2)

From Eq.2 it is possible to derive the state-transition probability

$$T(s', a, s) = p(s' | s, a) = P\{S_{t+1} = s' | S_t = s, A_t = a\} = \sum_{r \in R} p(s', r | s, a)$$

(3)

and the expected reward for state-action-next state triples:

$$r(s, a, s') = E[R_{t+1} | S_t = s, A_t = a, S_{t+1} = s'] = \frac{\sum_{r \in R} p(s', r | s, a)}{p(s' | s, a)}.$$

(4)

As described in Figure 3, at each time step t, the agent performs an action a based on the observed current system state. In the original MDP formulation, the state-action mapping is called policy and it is represented by a stochastic function,$\pi_t: A\ x\ S \rightarrow [0,1]$, where $\pi_t(s) = P(A_t = a | S_t = s)$ is the probability of performing the action a, being the system in the state s.

After every action, rewards are returned by the system and their values are higher if the performed actions are useful to reach the target system condition. It is intuitive that maximizing the rewards collected at each time

step t (i.e.,R_1, R_2, R_3 ...) leads to achieving the final goal on the system. This aspect is formalized through the discounted return, G_t, defined as:

$$G_t := R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{T} \gamma^k R_{t+k+1},$$

(5)

where the discounted rate $\gamma \in [0,1]$ determines the importance of future rewards. If γ is close to 0, the agent is myopic and maximizes only immediate reward; on the contrary, with, $\gamma \to 1$, agent becomes more farsighted.

Therefore, solving a MDP implies finding the optimal policy, π^*, maximizing G_t.

2.1.2. Policy and Value Functions

RL algorithms were developed to solve MDP providing an estimation of π^*. To this end, such methods rely on utility functions, called *state-value* function and *action-value* function. The state-value function, V^π: S $\to \mathbb{R}$ returns the expected value of G_t obtained by the agent by starting in the state s at time t and following the policy π thereafter:

$$V^\pi(s) := \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{T} \gamma^k R_{t+k+1}|S_t = s\right].$$

(6)

Similarly, the action-value function, or *Q-function*, Q^π: $S\ x\ A \to \mathbb{R}$ returns the expected value of G_t as consequence of performing action a when system state is s and then following the policy π:

$$Q^\pi(s,a) := \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{T} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right].$$

(7)

Both the value functions satisfy important recursive relationships, named Bellman equations [76]:

$$V^\pi(s) = \sum_{a\in A} \pi(a|s) \sum_{s'\in S, r\in R} p(s', r|s, a)[r + \gamma V^\pi(s')],$$

$$Q^{\pi}(s,a) = \sum_{s' \in S, r \in R} p(s', r|s,a)[r + \gamma V^{\pi}(s')] .$$

(8)

The state-value function defines a partial ordering relationship over policies, i.e., a policy $\pi' \geq \pi \leftrightarrow V^{\pi'}(s) \geq V^{\pi}(s) \; \forall \; s \in S$. Therefore, there is always at least one policy that is better than or equal to all the other policies, i.e., an optimal policy, π^*. Although there may be more than one optimal policy, they share the same state-value function, called optimal state-value function and defined as :

$$V^*(s) = \max_{\pi} V^{\pi}(s) \; \forall \; s \in S.$$

(9)

The optimal policy can, thus, be computed as:

$$\pi^*(s) = \arg \max_{\pi} V^{\pi}(s) \; \forall \; s \in S.$$

(10)

Optimal policies share also the same *optimal action-value function* Q^*:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \; \forall \; s \in S \text{ and } \forall \; a \in A .$$

(11)

As, the optimal action-value function $Q^*(s,a)$ and the optimal state-value function $V^*(s)$ are related by the equation:

$$V^*(s) = \max_{a} Q^*(s,a) \; \forall \; s \in S,$$

(12)

if $Q^*(s,a)$ is known, the optimal policy π^* can be derived by selecting action a^* which maximizes the optimal action-value function in each given state s, i.e.,

$$\pi^*(s) = a^* = \arg \max_{a} Q^*(s,a) \; \forall \; s \in S.$$

(13)

2.1.3. Q-Learning algorithm

Q-Learning (QL) is a RL algorithm that estimates the optimal policy directly from experience, i.e., by collecting sequences of transitions $< S_t, A_t, S_{t+1}, R_{t+1} >$, without the need of specifying the transition

probability function. QL is a value-based algorithm as it relies on a tabular approximation of the Q-function for each state-action pair. Given a set of N states and M actions, QL returns a matrix, $\boldsymbol{Q} \in \mathbb{R}^{N \times M}$, in which the element $Q[s_i, a_j]$ represents an approximation of the Q-function for the action a_j in state s_i. This means that the QL algorithm requires the state-action space to be discrete and small to produce accurate estimations. As illustrated in Figure 4, these characteristics allow to easily translate the $\boldsymbol{Q}$-matrix into a set of if-then-else rules. Indeed, by applying Eq.13 along its rows, it is possible to extract the optimal policy.

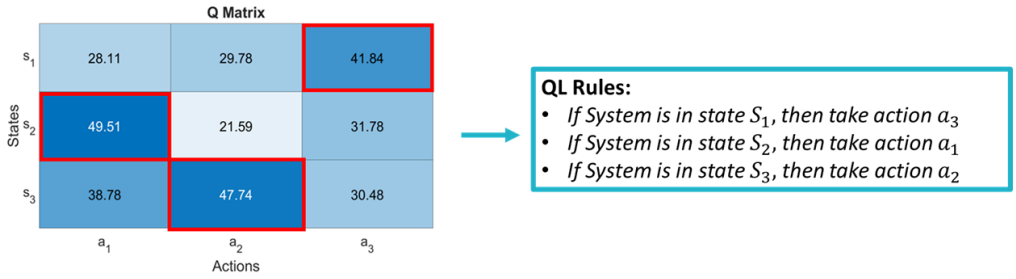


Figure 4: Translating into human readable "if-then-else" rules Q matrix optimal policy.

As reported in Algorithm 1, the learning of the $\boldsymbol{Q}$-matrix is an iterative process where a single episode composed by T steps is repeated for I times. Starting from an arbitrary matrix, at each time step t, the entry $Q[S_t, A_t]$ of the matrix is updated with:

$$\boldsymbol{Q}[S_t, A_t] \leftarrow \boldsymbol{Q}[S_t, A_t] + \alpha \left[R_{t+1} + \gamma \arg \max_{a} \boldsymbol{Q}[S_{t+1}, A_t] - \boldsymbol{Q}[S_t, A_t] \right]$$

(14)

where α is the *learning rate* that weighs the magnitude of the updates.

QL training is applied with a ε-greedy strategy to address the exploitation/exploration dilemma, i.e., the agent should not exploit only the currently estimated best policy but should also keep exploring the state-action space in order to potentially discover better strategies. When the ε-greedy strategy is adopted in QL training, at each time step t, the agent has a given probability ε of randomly taking an action instead of the best estimated action by using Eq. 13.

Algorithm 1: Pseudocode for Q-Learning algorithm.

Given: set of N states, set of M actions, learning rate α, discount factor γ, a probability ϵ, a maximum number of training iterations I, probability ϵ
init Q matrix arbitrarily
loop for each episode (I times):
    Set the current system state to S_0
    loop for each decisional time step t:
    p ← uniform random number ∈ [0,1]
    if p<ϵ
        Select action A_t randomly
    else
        $A_t \leftarrow arg\,max_a\, Q(S_t, a)$
    Perform A_t on the system
    Observe next state S_{t+1} and reward R_{t+1}
    $Q(S_t, A_t) \leftarrow$
        $Q(S_t, A_t) + \alpha \cdot [R_{t+1} + \gamma \cdot max_a\, Q(S_{t+1}, a) - Q(S_t, A_t)]$
  Set current system state to S_{t+1}

The easier interpretability of the estimated optimal policy and the low number of hyperparameters to tune, lead to the application of QL algorithm in the MIRL frameworks implemented in this thesis (section 2.2.).

2.2. Tackling a precision dosing problem with RL and PK-PD models

The aim of this section is to provide a detailed description of how RL and PK-PD models are coupled to solve precision dosing problems based on adaptive dosing strategies. Since RL algorithms were developed to solve a MDP, it is necessary to first introduce how such clinical workflow can be translated into an MDP.

Recalling the workflows of both adaptive dosing strategies (Figure 1) and MDP (Figure 3), their integration can be easily illustrated with the schematization in Figure 5.

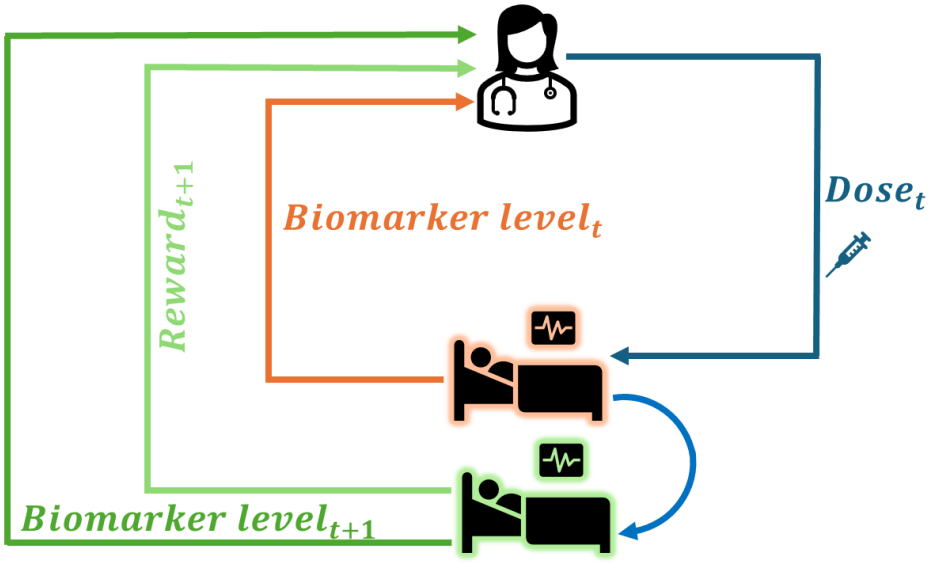


Figure 5: Mapping a Markov Decision Process to the context of adaptive dosing strategies based on patient monitoring.

In particular, the MDP agent/decision-maker corresponds to the clinician who periodically interacts with the system, which, instead, represents the patient. At each step of this interaction, the observation of the current system state is the clinician observing the actual patient health condition, for example by measuring the level of a biomarker related to treatment efficacy and/or safety. Then, in both the MDP and the precision dosing frameworks, the action-selection process is driven by the performed observation. At the next system/patient evaluation, both the RL agent and the clinician receive feedback (i.e., reward) on the performed action/dosing strategy by assessing whether the target system condition/therapeutic goal was achieved or not (e.g., keeping the biomarker in a target range).

Therefore, this formalization of adaptive dosing strategies in the MDP fashion allows RL-agent to optimize these precision dosing tasks and to learn the optimal dosing policy. To this end, as previously discussed, the RL agent would require a trial-and-error procedure based on trying different actions/doses at each system/patient state.

Obviously, this approach is practically unfeasible on actual patients due to safety and ethical concerns. A first solution partially circumventing this limitation requires the use of large clinical retrospective datasets to robustly estimate an optimal dosing strategy [30,39,71,77,78]. However, as underlined in section 1.3, such a large amount of data is rarely available, and when it is, these data sets often lack a wide variety of dosing scenarios, including those that could be particularly harmful to patient safety. Consequently, RL algorithms cannot be efficiently trained with only a data driven approach [39,73].

Recently, the integration of Pop-PK-PD models within the RL framework has been proposed to overcome the limitations due to the lack of training

data. This hybrid approach bridging Model-Informed Precision Dosing (section 1.2) and ML was called Model-Informed Reinforcement Learning (MIRL) and leverages the simulations of Pop-PK-PD models to generate the experience needed to robustly train RL-agents [30,39,72,73].

A first MIRL approach consists in using Pop-PK-PD models to generate synthetic datasets or, alternatively, to enlarge (i.e., data augmentation) the already available clinical information. In particular, Figure 6 provides a graphical description of the MIRL workflow leveraging Pop-PK-PD simulations to generate synthetic training sets for RL algorithms [39]. This method consists of two sequential steps. First, a large cohort of virtual patients, each characterized by a set of covariates and PK-PD parameters, is extracted. Then, different adaptive dosing scenarios are simulated on these patients by randomly selecting the administered drug amount at each decisional stage. This strategy aims to investigate the effect of different dosing protocols and leads to collect several transitions containing the current patient health status, the administered dose- and the subsequent evolution of patient conditions. In the second step, all the collected transitions are used to train the RL agent. At each training step, a transition is randomly extracted from the training set and the associated reward is computed and, finally, used to update the estimation of the RL optimal dosing strategy [39,73].

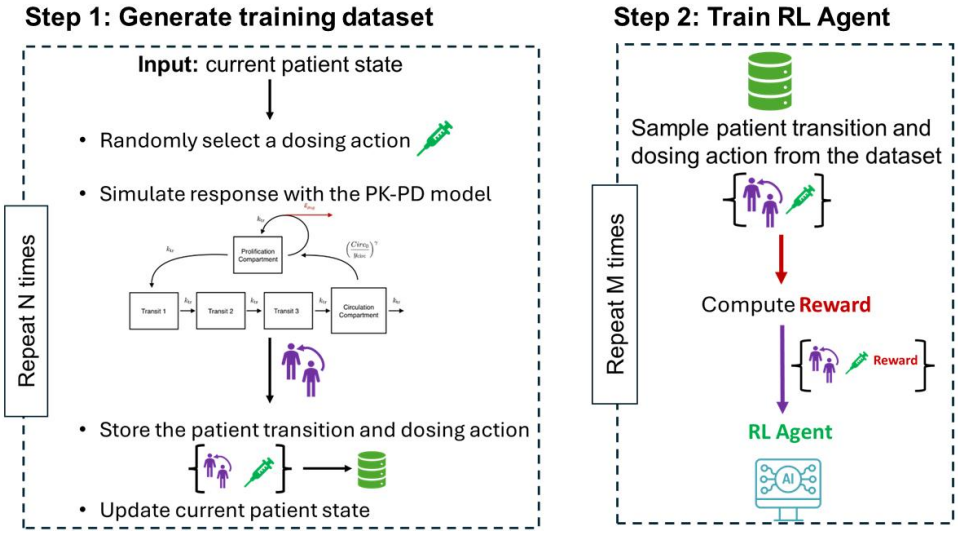


Figure 6: Schematical representation of a workflow integrating RL and PK-PD modelling to generate the experience necessary to train the RL agent for precision dosing tasks. In this case, two separated steps are required. First, PK-PD simulations are used to generate the training set by simulating a huge variety of dosing scenarios in the patients population. Dose selection is typically randomly performed. Each instance of the training set is a transition composed by the performed action, and both the current and the next states. Finally, once the synthetic dataset is obtained, it is used to train the RL agent to solve the precision dosing task (Step 2). At each training step, a transition is sampled from the training set, and it is used to estimate the optimal dosing policy of the RL agent.

Alternatively to the strategy in Figure 6, Pop-PK-PD models are embedded as simulation engines within the RL algorithms to predict the consequences of each dosing action. In particular, as illustrated in Figure 7, this workflow invokes the PK-PD model simulation following each dose adjustment selected by the RL algorithm. Thus, this integration perfectly replicates the agent-system interplay characterizing the MDP in which the RL algorithms dynamically and continuously learn and adapt the dosing strategies. Furthermore, it was demonstrated that this approach can facilitate the identification of optimal precision dosing regimens [73]. Consequently, the hybrid framework in Figure 7 was used as starting point of the methodologies implemented in this thesis.

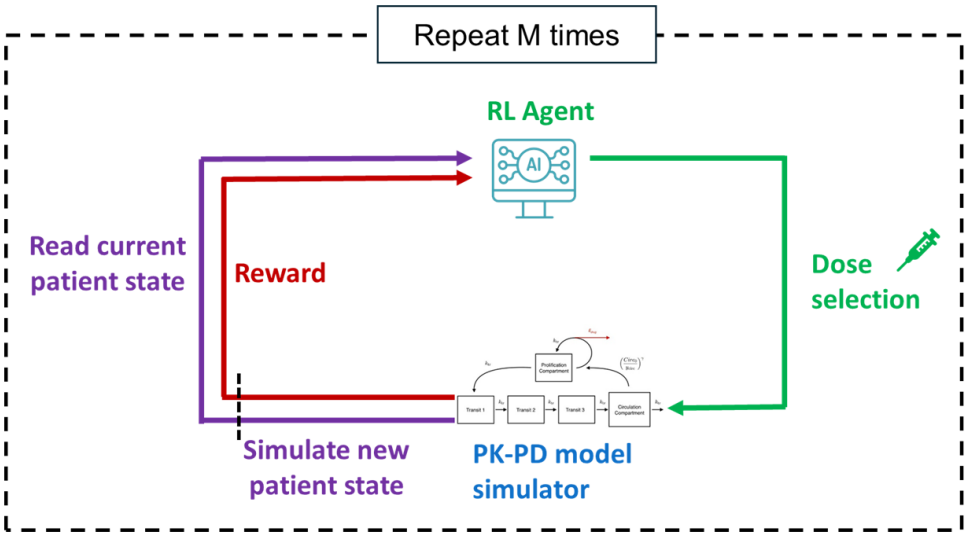


Figure 7: Hybrid workflow embedding PK-PD model simulation within the RL framework. At each decisional step, following the dosing strategy selected by the RL agent, a PK-PD model simulation is performed to simulate its effects. Then, the reward is computed and used by the RL agent to update its dosing policy.

As detailed in section 1.2, pharmacometrics models are leveraged in the precision dosing context to both optimize adaptive dosing protocols in target populations/sub-populations of patients and to tailor dose adjustments at single patient level by monitoring efficacy/toxicity biomarkers [27,45,49,62–66]. Therefore, in this thesis, starting from the workflow in Figure 7, different MIRL techniques were defined and evaluated to support both these two tasks.

More in details, section 2.2.1 will provide a description of the implemented framework to derive an optimal dosing protocol for an entire target population. Differently, in section 2.2.2 a novel MIRL approach to provide patient-tailored adaptive dosing rules will be presented. In both scenarios, all the developed methodologies leverage QL as RL algorithm due

to its easy interpretability of the estimated optimal dosing policy (Figure 4) and the low number of parameters to fine-tune (Algorithm 1). However, the following approaches can be easily extended to the other RL techniques reported in Appendix A.

2.2.1. Learning an optimal adaptive dosing protocol for an entire patient population with RL and PK-PD models

Optimizing an adaptive dosing protocol on a target population implies finding general dose-adjustment rules that satisfy efficacy and toxicity criteria for most of the individuals. From a MIRL perspective, this aspect leads to the use of a unique RL agent to guide the treatment of all the patients. Therefore, considering IIV of the PK-PD response during the training of the RL algorithm is essential to obtain generalizable adaptive-dosing strategies able to balance the efficacy/toxicity tradeoff in the target population. To this end, starting from the seminal works available in literature [39,71–74,78], the workflow in Figure 7 was adapted in this thesis as illustrated in Figure 8.

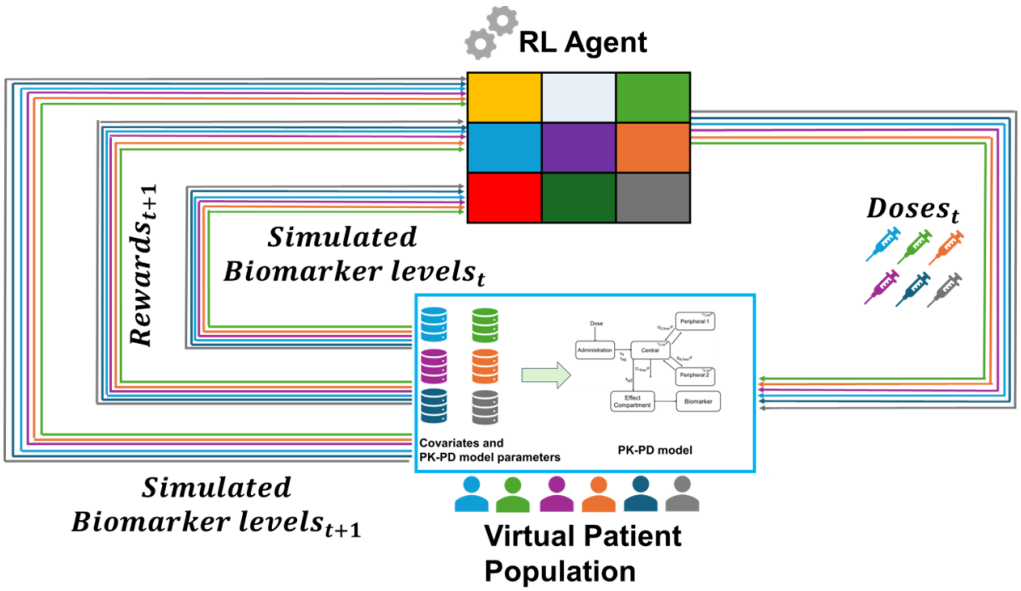


Figure 8: Model-Informed Reinforcement Learning workflow to estimate a general adaptive dosing protocol for an entire population of patients. RL agent was represented in a tabular form coherently with the Q-Learning algorithm used in this thesis.

In particular, a single RL agent is trained to simultaneously optimize the drug administration in all the patients (i.e., in terms of covariates and PK-PD model parameters) of the target population. Therefore, at each time step, PK-PD simulations outputs are fed to the agent in order to assess the current health status of the patients. Then, for each of them, a specific dose is selected and, its effect (i.e., next state) is assessed by running other PK-PD

simulations. Finally, the rewards in the population are computed and integrated to update the estimation of the RL-based optimal dosing policy.

Merging the reward coming from the different individuals is central to derive a general and robust, with respect to IIV, dosing protocol for the whole patient population. Indeed, through this procedure, the RL agent can evaluate whether the same dose administered in different patients (e.g., with different disease/PK-PD characteristics) having the same health status (e.g., observed biomarker level), is generally suitable to optimize the efficacy/safety balance in the population.

In the implemented framework, updates of the optimal policy coming from different patients but referred to the same state-action pair, were merged by performing an averaging operation. This strategy was preferred to techniques reported in literature as led to a stable estimation of the action-value function (Eq. 8) and avoided the introduction of further layers of complexity (e.g., hyperparameters) in the RL algorithm [76,96,97].

Algorithm 2 summarizes the framework in Figure 8 through a pseudocode. Its formulation is based on QL (section 2.1.3) as it was used in all the case studies presented in this thesis due to its easy interpretation of the estimated optimal policy and the lower number of hyperparameters (section 2.1.3.). Consequently, a QL-specific updating strategy was defined to merge the contributions coming from the different patients for each state-action couple. However, the following considerations can be easily extended to the other RL algorithms presented in Appendix A.

As reported in Algorithm 2, at each decisional step, for the $l-th$ patient in the target population, the transition, $b =< s_{i,l}, a_{j,l}, s'_{m,l}, r_l >$, is stored in the buffer B. Note that subscripts i and m refer to a system/patient state belonging to the finite N-dimensional set S; j to an action within the M-dimensional set A. Then, recalling Eq.14, for each s_i, a_j couple in B, the average update $u(s_i, a_j)$ is computed through Eq.15

$$u(s_i, a_j) = \frac{1}{K}\sum_{k=1}^{K} r_k + \gamma \cdot \max_a Q(s'_{m,k}, a),$$

(15)

with K being the number of transitions in B containing s_i, a_j. Finally, each $Q(s_i, a_j)$ is updated with:

$$Q(s_i, a_j) = Q(s_i, a_j) + \alpha \cdot [u(s_i, a_j) - Q(s_i, a_j)].$$

(16)

Algorithm 2: Pseudocode of the framework integrating PK-PD model simulations and QL algorithm to estimate an optimal general dosing protocol.

Given: population of L patients each with a set of covariates set $\boldsymbol{X_l}$ and of PK-PD parameters $\boldsymbol{\theta_l}$, PK-PD model $f(t, \boldsymbol{X}, \boldsymbol{\theta})$, QL algorithm, O number of treatment cycles, I number of training iterations, B transition buffer, D update buffer, probability ϵ
init QL agent
loop for each training iteration (I times):
  loop for each patient in the population (L times):
    Set patient initial state
  loop for each decisional step (O times):
    loop for each patient in the population (L times):
      QL agent selects a_j level for the $l-th$ patient considering the current patient state s_i
      simulate the effect by calling $f(t, \boldsymbol{X_l}, \boldsymbol{\theta_l})$
      update new patient health status s'_m
      compute reward r
      store the transition $b =< s_{i,l}, a_{j,l}, s'_{m,l}, r_l >$ in B
    loop for each s_i, a_j couple in B:
      compute the average QL update $u(s_i, a_j)$ using Eq.15
      update QL agent decisional policy with Eq. 16

2.2.2. Learning patient-specific adaptive dosing strategies with RL and PK-PD models

In the context of precision dosing based on adaptive dosing strategies, pharmacotherapies can be customized at a single individual level rather than on the entire population/sub-population. When the focus is on the single subject, the MIRL approach presented in section 2.2.1 does not properly fit to this patient-centric optimization. Indeed, such method estimates the optimal dosing policy by considering the effects of a dosing strategy on the entire patient population (i.e., integrating the rewards returned by all individuals). Furthermore, it was shown that in presence of high IIV (especially when not explainable by the covariates), protocols based on the MIRL method in section 2.2.1 can work sub-optimally or inefficiently for some patients if their PK-PD characteristics are strongly deviated by the typical ones [79,98].

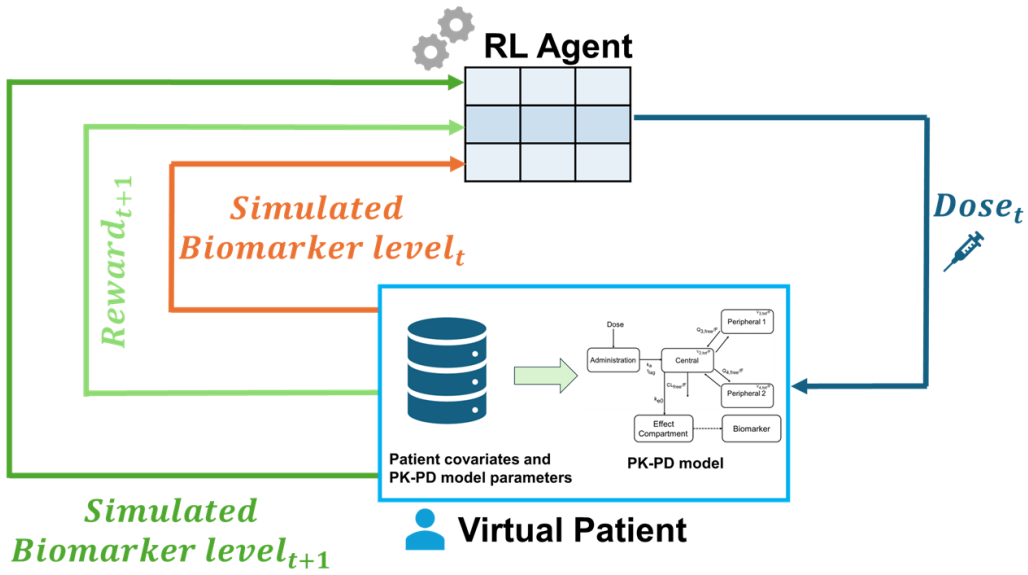


Figure 9: Schematization of the novel Model Informed Reinforcement Learning framework introduced to derive patient-specific adaptive dosing protocols. For each individual a personal RL agent is used to customize the dosing schedule. To this end, a virtual patient (i.e., a PK-PD model with patient covariates and parameters) is used to generate the experience necessary to learn the optimal RL-based dosing protocol. RL agent is represented in the tabular form of QL as it represented the main RL algorithm investigated in this thesis.

To address these issues, following the seminal work of Meier et al. (2021) [80], a novel MIRL approach is here introduced to tailor an adaptive dosing strategy on a single patient (Figure 9).

In this new paradigm, the shift in treatment optimization from an entire population to a specific individual translates to moving from a unique RL agent trained on an entire patient population, to a personal RL agent trained on a specific patient. Therefore, each patient in the population will be treated following the adaptive dosing protocol obtained by a dedicated patient-specific RL agent.

Another central feature of this framework is the leveraging of patient digital twin to provide the experience necessary to train the RL agent for solving the precision dosing task [60,61]. This virtual replica of the patient (i.e., virtual patient in the schematization of Figure 9), is a PK-PD model with an individual set of parameters and covariates describing the pharmacological response for that specific individual.

Assuming that the virtual patient adequately represents the real patient, PK-PD simulations are embedded within the RL framework to estimate the personalized optimal dosing policy. In particular, as illustrated in Figure 9, the patient virtual twin is used to mimic the agent-system trial-and-error interplay typical of MDP. Therefore, the RL agent can dynamically experiment and evaluate through the reward signals several dosing strategies

on the virtual patient. Thus, the optimal adaptive-dosing strategy for the single patient can be learnt following an exhaustive exploration of the decisional space.

Obviously, this framework is not feasible without using a virtual twin, as it would require testing several dosing strategies directly on the patient and then, use this data to train the RL agent.

The implementation of the novel workflow in Figure 9 is formally described with the pseudocode in Algorithm 3. Although its conceptualization is reported for Q-Learning algorithm (section 2.1.3), it can be easily generalized and extended to the other RL algorithms presented in Appendix A.

Algorithm 3: Pseudocode of the novel MIRL framework to tailor adaptive dosing strategy on a single patient.

Given: patient with a set of covariates $\boldsymbol{X_l}$ and of PK-PD parameters $\boldsymbol{\theta_l}$, PK-PD model $f(t, \boldsymbol{X}, \boldsymbol{\theta})$, QL algorithm, O number of treatment cycles, I number of training iterations, probability ϵ
init QL agent
loop for each training iteration (I times):
    Set initial patient state
    loop for each decisional step (O times):
        QL agent selects a_j level for the $l-th$ patient considering the current patient state s_i
        simulate the effect of a_j by calling $f(t, \boldsymbol{X_l}, \boldsymbol{\theta_l})$
        update new patient health status s'_m
        compute reward r
        update QL agent

Chapter 3

Optimizing a single efficacy/toxicity endpoint. Application of the RL/PK-PD framework to the erdafitinib precision dosing problem[2]

This chapter aims to present a first application of the novel MIRL paradigm described in section 2.2 to tailor adaptive dosing strategies on each specific patient.

Here, it is evaluated on a relevant case study directly derived from clinical oncology, i.e., the erdafitinib (Balversa, Janssen Pharmaceutical companies) precision dosing in metastatic urothelial cancer patients [99–103]. According to FDA guidelines, erdafitinib administration follows an adaptive dosing protocol based on the monitoring of the serum phosphate concentration levels ($[PO_4]_{serum}$), acting as efficacy/safety biomarker [104]. However, due to the narrow therapeutic window and significant IIV of pharmacological response [100], erdafitinib therapy could potentially benefit from a personalization of the adaptive dosing protocol at individual patient level.

Therefore, the novel MIRL framework described in section 2.2.2 is used to customize the adaptive dosing rules of erdafitinib on each patient deriving clinically acceptable administration strategies. To provide a robust evaluation of this patient-centric MIRL approach, its performances are benchmarked by the FDA approved erdafitinib adaptive dosing rules and the MIRL-based general protocol obtained with the methodology illustrated in section 2.2.1.

This chapter is structured as follows. First, erdafitinib precision dosing problem and its FDA approved adaptive dosing protocol will be illustrated in section 3.1.1. Then, sections 3.1.2.1.-2-3.1.2.5. will show how to setup the

[2] The content of this chapter is published in [79].

MIRL framework to optimize erdafitinib precision dosing. Consequently, the focus will be on showing the formalization of this case study using the essential bricks of a RL algorithm (QL). Sections 3.1.2.6. and 3.1.2.7. will describe the technical implementation of the MIRL frameworks and the evaluation study performed, respectively. Finally, sections 3.2 and 3.3 will illustrate and discuss the obtained results. Supplementary materials supporting this chapter are reported in Appendix B.

3.1. Methods

3.1.1. Erdafitinib case-study

Erdafitinib is an orally administered inhibitor of the fibroblast growth factor receptors (FGFR) approved for the treatment of metastatic urothelial carcinoma [99–103]. $[PO_4]_{serum}$ level is the most important efficacy/safety biomarker for erdafitinib treatment [99,100,105]. Indeed, $[PO_4]_{serum}$ increases as a consequence of the FGFR target engagement [102,106] and a $[PO_4]_{serum}$ level >5.5 mg/dL was found to correlate with a better response in terms of progression free survival [99]. At the same time, hyperphosphatemia, i.e., an excessive $[PO_4]_{serum}$ level ($\geq 7mg/dL$) is the major cause of drug-related adverse events [99].

Consequently, the FDA-approved erdafitinib protocol is based on an adaptive dosing strategy (Figure 10) aiming to maintain the $[PO_4]_{serum}$ within the target range, $[5.5, 7)\ mg/dL$ [104]. Briefly, five discrete dose levels, i.e., 4, 5, 6, 8 and 9 mg/day, are available. Treatment always starts at a dose of 8 mg/day. After two weeks (14th day) of treatment patients are monitored and initial dose increased to 9 mg/day in case of insufficient efficacy, i.e., $[PO_4]_{serum} < 5.5\ mg/dL$. After the first evaluation, patient monitoring typically occurs on a monthly basis. In case of hyperphosphatemia events, erdafitinib administration is temporarily interrupted, patient are monitored weekly, and treatment is resumed when $[PO_4]_{serum}$ returns $< 5.5\ mg/dL$ (Figure 10, Panel A). The resumption dose depends on the hyperphosphatemia severity, discretized in three grades, i.e., $[7, 9]\ mg/dL$, $(9,10]\ mg/dL$ and $> 10\ mg/dL$, and on the duration of the treatment interruption (Figure 10, Panel B).

A population PK-PD model describing the erdafitinib PK and its effect on $[PO_4]_{serum}$ is available [99–101]. Model equations and population parameter values are reported in Appendix B. Based on this PK-PD model, an *in silico* evaluation of the FDA-approved adaptive dosing protocol of erdafitinib was performed [100]. It revealed that, after four months of treatment (i.e., the median time for assessing treatment efficacy [100]), the 30% of patients was expected to have $[PO_4]_{serum}$ within the target range, $[5.5, 7)\ mg/dL$. It is also highlighted a high IIV of pharmacological response [100].

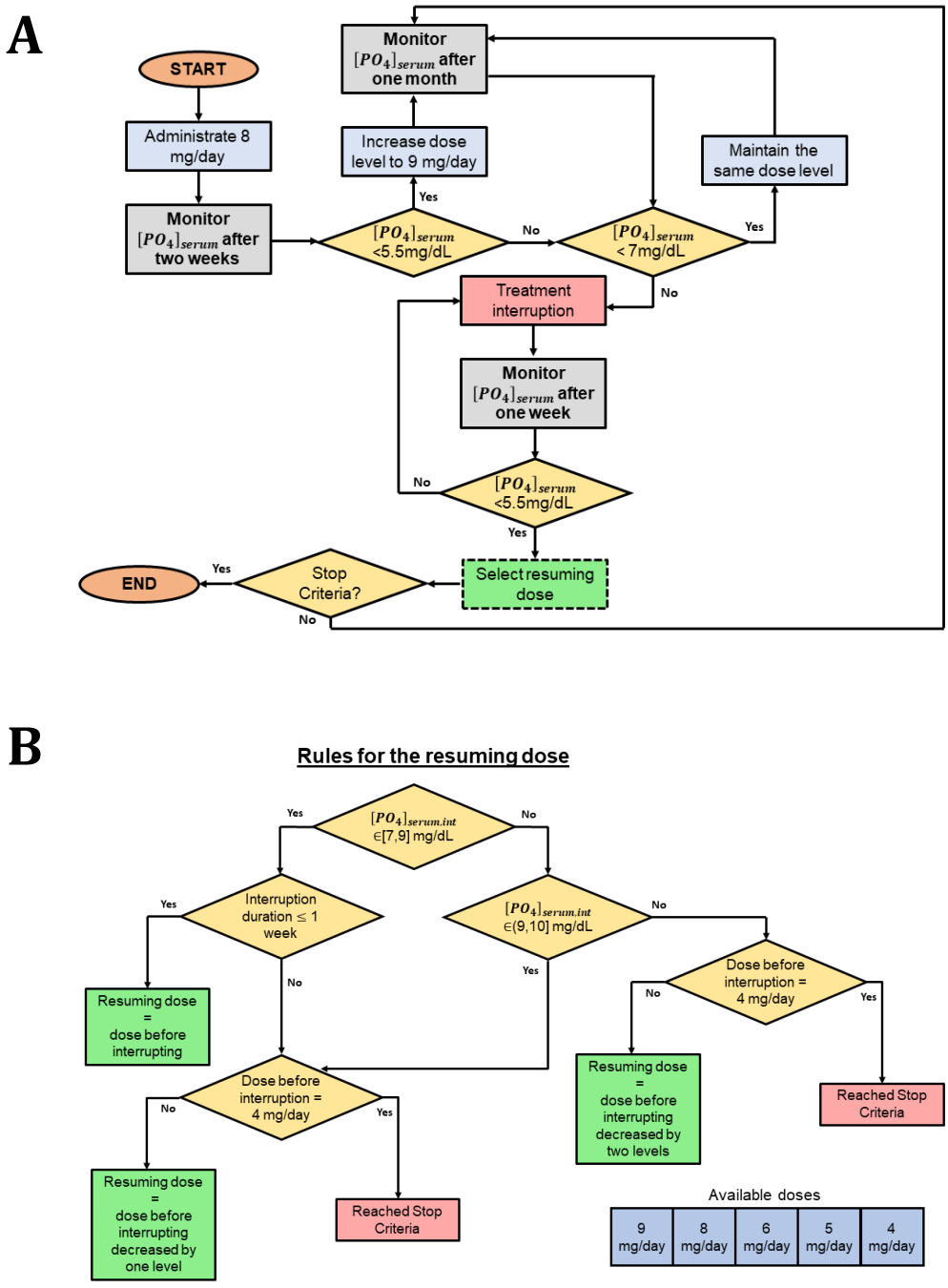


Figure 10: Flow chart of the FDA approved adaptive-dosing protocol of erdafitinib. Dose levels are adjusted according to the $[PO_4]_{serum}$ level (Panel A). In case of treatment interruption, the resuming dose is selected based on the severity of the hyperphosphatemia event and on the interruption duration (Panel B).

3.1.2. RL-based precision dosing of erdafitinib

Erdafitinib precision dosing problem was formalized through a finite MDP (section 2.1.1.). Therefore, the three key elements of a MDP, i.e., system states, agent actions and reward function, have to be designed

coherently with the clinical application. System state, i.e., patient status, is described by $[PO_4]_{serum}$ which is periodically monitored to guide dose adjustment (i.e., agent action). As the clinical goal is keeping $[PO_4]_{serum}$ within the target range $[5.5,7)\ mg/dL$, thus guaranteeing a phosphate level above the efficacy threshold and simultaneously avoiding hyperphosphatemia events, dose adjustments are evaluated by a reward function depending on the $[PO_4]_{serum}$ level. QL was used as RL algorithm in the optimization of erdafitinib precision dosing problem. In particular, QL was integrated with patient digital twin/virtual population depending on the MIRL framework and the level of treatment personalization (sections 2.2.1. and 2.2.2.). To this end, the erdafitinib PK-PD model (section B.1. of Appendix B) linking drug doses with $[PO_4]_{serum}$ and individual characteristics (i.e., both patient's covariates and PK-PD parameters) were leveraged. A more detailed description of the RL-based formalization of erdafitinib precision dosing problem is discussed in sections 3.1.2.1-3.1.2.3.

3.1.2.1. Reward function

The reward function was a $\mathbb{R}^D \rightarrow \mathbb{R}$ function of the $\mathbb{R}^D$ vector, $[\boldsymbol{PO_4}]_{\boldsymbol{obs}} = ([PO_4]_{obs,1}, \dots [PO_4]_{obs,D})$, containing D observations of daily $[PO_4]_{serum}$, where D depends on the number of days between two consecutive patient monitoring, i.e., 7, 14 or 28 days. The reward is given by the sum of two contributes weighted by the coefficients β_1 and β_2 (Eq.17):

$$Reward([\boldsymbol{PO_4}]_{\boldsymbol{obs}}) = \beta_1 \cdot \sum_{i=1}^{D} g\left([PO_4]_{obs,i}\right) + \beta_2 \cdot \frac{\sum_{i=1}^{D} h([PO_4]_{obs,i})}{D}$$

(17)

where the expressions of the functions g and h are reported in Eq.18 and Eq. 19, respectively.

$$g\left([PO_4]_{obs,i}\right) = \begin{cases} 0.5 \cdot \exp\left(-\lambda_2 \cdot \left[-\left([PO_4]_{obs,i} - 5.5\right)\right]\right) & if\ [PO_4]_{obs,i} < 5.5\ mg/dL \\ \min\left(1, \left([PO_4]_{obs,i} - 5.5\right) + 0.5\right) & if\ [PO_4]_{obs,i} \in [5.5,6.25]\ mg/dL \\ \min\left(1, \left([PO_4]_{osb,i} - 6.5\right) + 1\right) & if\ [PO_4]_{obs,i} \in (6.25,7)\ mg/dL \\ 0.5 \cdot \exp\left(-\lambda_1 \cdot \left([PO_4]_{obs,i} - 7\right)\right) & if\ [PO_4]_{obs,i} \in [7,9]\ mg/dL \\ 0.18 \cdot \exp\left(-\lambda_2 \cdot \left([PO_4]_{obs,i} - 9\right)\right) & if\ [PO_4]_{obs,i} > 9\ mg/dL \\ \lambda_1 = 0.5, \quad \lambda_2 = 1.5 & \end{cases}$$

(18)

$$h\left([PO_4]_{obs,i}\right) = \begin{cases} 1 & [PO_4]_{obs,i} \in [5.5,7)\ mg/dL \\ 0 & otherwise \end{cases}$$

(19)

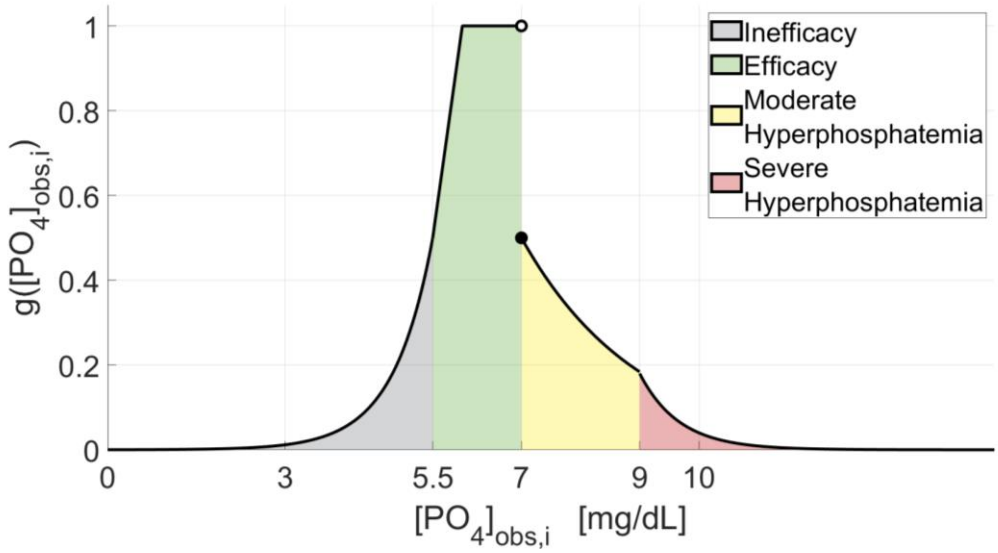


Figure 11: Reward assigned to daily observed $[PO_4]_{serum}$, $[PO_4]_{obs,i}$, with function g (Eq.18).

The first term in Eq.17 accounts for the $[\mathrm{PO_4}]_{\mathrm{serum}}$ levels as transformed by function g. As illustrated in Figure 11, it assigns higher reward to actions bringing $[\mathrm{PO_4}]_{\mathrm{serum}}$ close to the efficacy range with the maximum renumeration for the $[6.25,7)$ mg/dL interval since a higher $[\mathrm{PO_4}]_{\mathrm{serum}}$ is associated with a better response in terms of overall survival and progression free survival [99]. Conversely, actions leading to hyperphosphatemia or inefficacy are penalized. The penalty becomes more significant as the biomarker moves further away from the target window. The second term assigns an extra reward proportional to the percentage of days in a treatment cycle in which $[\mathrm{PO_4}]_{\mathrm{serum}}$ is maintained in the $[5.5,7)$ mg/dL range. The coefficients β_1 and β_2 were arbitrary fixed to 10 and 5, respectively. In this way, the first term in Eq. 17, that represents the core of the reward function, has an higher weight than the second one, that was introduced to stress the importance of maintaining $[\mathrm{PO_4}]_{\mathrm{serum}}$ in the target range and to increase the gap, in terms of reward, between actions leading to efficacy and those bringing $[\mathrm{PO_4}]_{\mathrm{serum}}$ to toxicity or inefficacy.

3.1.2.2. System/Patient health state

The state of the system (patient) is described by a tuple of three elements, $X = \{L_{PO_4}, D_P, L_{int}\}$. The first, $L_{PO_4,}$ accounts for the $[PO_4]_{serum}$ observation at the monitored day, discretized in six levels, as detailed in Eq.20:

$$L_{PO_4}([PO_4]_{serum}) = \begin{cases} 1 \quad if\ [PO_4]_{serum} < 5.5\ mg/dL \quad (inefficacy) \\ 2 \quad if\ [PO_4]_{serum} \in [5.5,6)\ mg/dL \quad (efficacy-1) \\ 3 \quad if\ [PO_4]_{serum} \in [6,6.5)\ mg/dL\ (efficacy-2) \\ 4 \quad if\ [PO_4]_{serum} \in [6.5,7)\ mg/dL\ (efficacy-3) \\ 5 \quad if\ [PO_4]_{serum} \in [7,9]\ mg/dL \quad (moderate\ hyperphosphatemia) \\ 6 \quad if\ [PO_4]_{serum} > 9\ mg/dL \quad (severe\ hyperphosphatemia). \end{cases}$$

(20)

In particular, the efficacy range $[5.5,7)$ mg/dL was subdivided into three intervals to allow the QL agents to select different actions depending on whether the $[PO_4]_{serum}$ is located in the lower ($[5.5,6)$ mg/dL), the middle ($[6, 6.5)$ mg/dL) or the upper ($[6.5,7)$ mg/dL) part of the target window. The second element of patient state, D_P, codes for the last administered dose. Only the five clinically available dose levels, i.e., {4,5,6,8,9} mg/day, were considered. Dose level is coupled with a sign (+/-) that is negative if the treatment is temporarily interrupted, positive otherwise (Eq. 21). This was an arbitrary convention adopted to store two pieces of information, i.e., the temporary interruption and the triggering dose, in a compact form. Indeed, storing information on the dose leading to an interruption is necessary to choose the resumption dose. A flag value equal to -1 encodes for the treatment beginning in which the initial dose must be determined.

$$D_P = \begin{cases} -1 & for\ the\ initial\ state \\ -4, -5, -6, -8, -9 & for\ doses\ provoking\ a\ temporary\ treatment\ interruption \\ 4\ ,5, 6, 8, 9 & otherwise. \end{cases}$$

(21)

Finally, L_{int} stores the discretized $[PO_4]_{serum}$ level (Eq.22) that caused a temporary treatment interruption, $[PO_4]_{serum,int.}$. As the severity of hyperphosphatemia event triggering treatment interruption is a key clinical factor for the choice of restarting the treatment and of the resuming dose, L_{int} was defined to distinguish between severe ($L_{int} = 6$) and moderate ($L_{int} = 5$) conditions. A flag value -1 was arbitrary assigned to states not associated with a temporary interruption.

$$L_{int}([PO_4]_{serum}) = \begin{cases} 5 \quad if\ [PO_4]_{serum,int.} \in [7,9]\ mg/dL \\ 6 \quad if\ [PO_4]_{serum,int.} > 9\ mg/dL \\ -1 \quad no\ interruption \end{cases}$$

(22)

An initial set of 60 couples $\{L_{PO_4}, D_P\}$ was defined by combining all the L_{PO_4} values with each $D_P \neq -1$. Couples associated with treatment interruption, i.e., with $D_P < -1$, were combined with all the positive values of L_{int}; conversely, the ones with $D_P > 0$ were combined only with $L_{int} = -1$, obtaining 90 different states. Since the temporary treatment interruption causes a $[PO_4]_{serum}$ decrease, interruption states with a post-discontinuation

$[PO_4]_{serum}$ greater than the before-discontinuation $[PO_4]_{serum}$ (i.e., $L_{PO_4} > L_{int}$) were not considered. Thus, the tuples with $D_P < -1$, $L_{PO_4} = 6$ and $L_{int} = 5$ were removed, reducing the state number to 85. Since at the beginning of the treatment all patients have $[PO_4]_{serum} < 5.5\ mg/dL$ [100], only one initial state with $L_{PO_4} = 1$, i.e. $X_I = \{1, -1, -1\}$, was added to the states space, thus leading the final number of states to 86.

3.1.2.3. Agent Actions

Actions were designed to give the agent more freedom of choice with respect to the clinical protocol rules in order to allow the identification of potentially better personalized treatments. However, safety constraints derived from the clinical practice, such as gradual dose variation, treatment interruption due to toxicity or dosage reduction post-interruption, were considered. Table 1 summarizes all the possible agent actions stratified by system states. Differently from the clinical protocol (Figure 10), the agent can select the initial dose among all the possible dose levels and decide to increase ($D+$)/decrease ($D-$) by one level or maintain ($D=$) the previous administered dose at each patient monitoring step, in absence of hyperphosphatemia. Adverse events are differently tackled by the agent according to their severity. In case of moderate hyperphosphatemia ($L_{PO_4} = 3$), agent can decide to temporarily interrupt the treatment (Int) or to continue erdafitinib administration with a dose lower or equal to the previous one, D_P. In case of severe hyperphosphatemia ($L_{PO_4} = 4$), the agent is constrained to interrupt the treatment. After that, treatment can be resumed with the dosage-rules reported in Table 1 only if $L_{PO_4} < L_{int}$, otherwise interruption is maintained. Conversely, clinical protocol enables treatment resumption only if $[PO_4]_{serum} < 5.5\ mg/dL$ ($L_{PO_4} = 1$).

Table 1: Summary of the agent actions stratified by each system/patient state.

D_P	Allowed Actions
Initial State	
-1	$4, 5, 6, 8, 9\ [mg/day]$
States without hyperphosphatemia: $L_{PO_4} \leq 4$ and $L_{int} = -1$	
4	$D+, D=$
5	$D+, D-, D=$
6	$D+, D-, D=$
8	$D+, D-, D=$
9	$D-, D=$
States with moderate hyperphosphatemia: $L_{PO_4} = 5$ and $L_{int} = -1$	
4	$Int., 4\ [mg/day]$
5	$Int., 4, 5\ [mg/day]$
6	$Int., 4, 5, 6\ [mg/day]$
8	$Int., D=, , 4, 5, 6, 8\ [mg/day]$

9	$Int., 4, 5, 6, 8, 9\ [mg/day]$
States with severe hyperphosphatemia: $L_{PO_4} = 6$ and $L_{int} = -1$	
For each D_P	$Int.$
States in which treatment can be resumed: $L_{PO_4} < L_{int}$ and $L_{int} > 0$	
-4	$0, 4\ [mg/day]$
-5	$0, 4, 5\ [mg/day]$
-6	$0, 4, 5, 6\ [mg/day]$
-8	$0, 4, 5, 6, 8\ [mg/day]$
-9	$0, 4, 5, 6, 8, 9\ [mg/day]$
States in which treatment cannot be resumed: $L_{PO_4} = L_{int}$ and $L_{int} > 0$	
For each D_P	$0\ mg/day$

3.1.2.4. Implementation

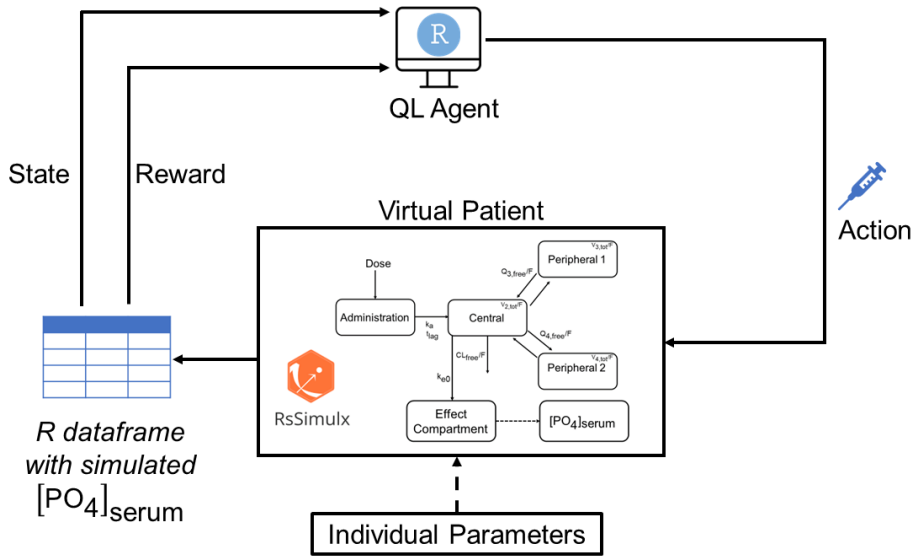


Figure 12: Implementation of the Model-Informed Reinforcement Learning framework to tackle erdafitinib precision dosing problem. In this case, two different software were integrated. R was used to code Q-Learning algorithm while Simulx was leveraged to simulate erdafitinib PK-PD response.

Figure 12 illustrates the implemented workflow for the MIRL approaches described in Chapter 2. In particular, a simulation platform iteratively predicting the response to adaptive dosing administration of erdafitinib was embedded within the QL-framework to allow the continuous interaction between the QL-agent and the virtual patient/population. The algorithm was coded in R Version 3.4.1 (Comprehensive R Network, http://cran.r-project.org/). The simulation platform was developed based on the erdafitinib PK-PD model [100,101] (section B.1. of Appendix B), coded in Mlxtran and simulated through the R package RsSimulx (Simulx 2021R1, Lixoft SAS, a Simulations Plus company). At each treatment cycle, the QL-

agent decided the dose level for the next cycle on the basis of the patient status, the $[PO_4]_{serum}$ response was simulated by the model and the simulated output used to update the patient status and compute the reward. From a methodological perspective, this hybrid framework coupling R with Simulx represents an absolute novelty in the MIRL landscape [39].

The same workflow was used for simulating the FDA-approved protocol. In this case, the decisional engine is the adaptive dosing rules reported in Figure 10. All simulations were performed without considering residual unexplained variability (RUV).

3.1.2.5. Evaluation setup

The MIRL frameworks presented in sections 2.2.1. and 2.2.2. were integrated with QL algorithm and, then, challenged to personalize a five-months treatment (i.e., average duration of therapy [107]) in a heterogeneous population of 141 virtual patients that was generated by the stratified random sampling strategy described in section B.2. of Appendix B. Individual parameters and covariates of virtual patients were extracted from erdafitinib population PK/PD model and the covariates distribution of subjects on which the model was estimated (sections B.1. and B.2. of Appendix B). As detailed in section B.2. of Appendix B, a stratified sampling was applied to obtain a heterogeneous virtual population in terms of treatment response. This strategy is crucial to guarantee a robust evaluation as it allows to explore the widest possible range of treatment responses.

In particular, the virtual population includes both completely responsive subjects (n=96), i.e., subjects for which $[PO_4]_{serum}$ level can be within the target range at treatment end, and partially (n=45) responsive subjects, i.e., subjects for which erdafitinib becomes no longer effective before the fifth month, (see details in section B.2.3. of Appendix B).

In order to obtain personalized adaptive dosing strategies with the MIRL framework in section 2.2.2., an individual QL-agent (named QLind-agent) was trained for each virtual patient (algorithm hyperparameters in Table B.8). The individual QL-based protocols were compared with the FDA-approved protocol (Figure 10) and with a population QL-based protocol learned by training a single QL-agent for the entire population (QLpop-agent) of patients.

3.2. Results

The focus of this section will be on presenting the performances of the personalized QL-based adaptive dosing protocols. In particular, section 3.2.1. will show the comparison of this MIRL approach with the FDA-approved dosing rules for erdafitinib. Differently, in section 3.2.2 the attention will be moved to analyze the differences and the analogies between

the results obtained by QL-based personalized dosing protocols and the general QL-based population protocol.

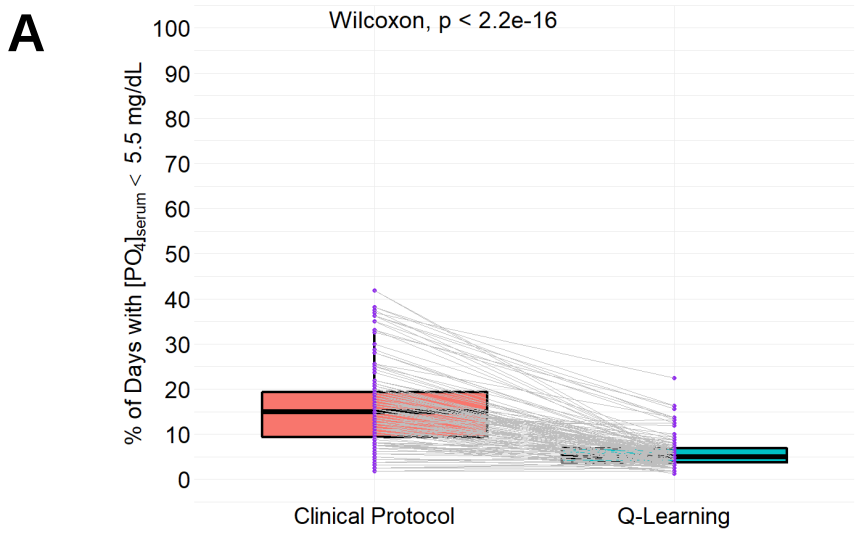
3.2.1. RL-based personalized dosing strategies vs FDA-approved protocol

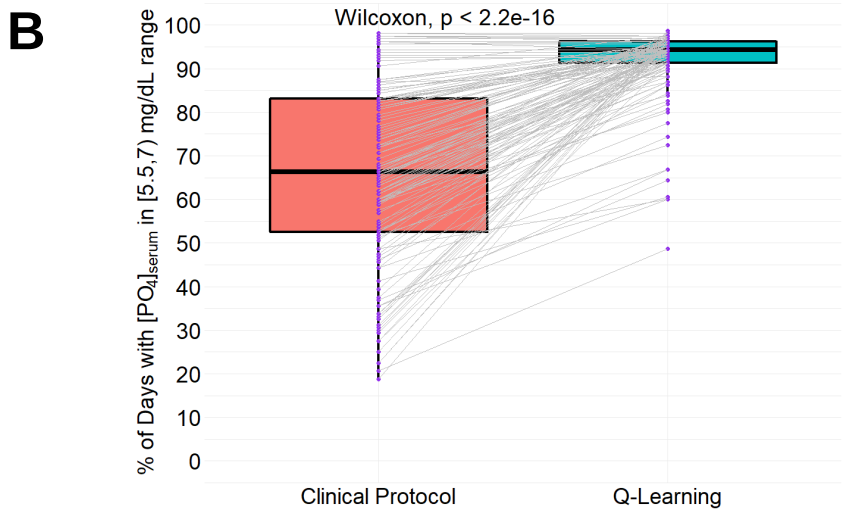
For each subject of the virtual population (n=141), an individual QL agent (QLind-agent) was trained to personalize the erdafitinib adaptive dosing protocol on a time-window of five months. To obtain a valid evaluation benchmark for the QL-based individual protocols, a five-months treatment was simulated on the same virtual population following the FDA-approved adaptive dosing rules. As reported in Table 2, the two strategies were compared in terms of percentage of patients with $[PO_4]_{serum}$ in the ranges of inefficacy ($< 5.5\, mg/dL$), efficacy ($[5.5,7)\, mg/dL$) , moderate ($[7,9]\, mg/dL$) and severe ($> 9\, mg/dL$) hyperphosphatemia. Three clinically relevant time-points were considered: the second week of treatment, i.e., first patient evaluation, the end of fourth month, i.e., median time for assessing erdafitinib efficacy [99,100], and the end of treatment.

Table 2: Comparison between the clinical and QL-based individual protocols at time-points of interest. Percentages of patients with $[PO_4]_{serum}$ within a certain range were used as evaluation metric. Results are reported for both the overall population and the subgroup of responsive patients.

Range of $[PO_4]_{serum}$ $[mg\ /dL]$	**%patients in the whole population [141 subjects]**		**%patients in the subgroup of completely responsive patients [96 subjects]**	
	Clinical Protocol	**QL-based individual protocols**	**Clinical Protocol**	**QL-based individual protocols**
Evaluation at the second week of treatment				
< 5.5	1.42	2.13	2.08	3.13
[5.5,7)	34.05	76.59	31.25	77.08
[7,9]	48.93	21.28	50.00	19.79
> 9	15.60	0.00	16.67	0.00
Evaluation at the end of the fourth month of treatment				
< 5.5	17.73	0.00	6.25	0.00
[5.5,7)	70.21	97.87	76.05	96.87
[7,9]	12.06	2.13	17.70	3.13
> 9	0.00	0.00	0.00	0.00
Evaluation at the end of treatment (fifth month)				
< 5.5	36.88	31.91	7.29	0.00
[5.5,7)	56.73	68.09	83.33	100
[7,9]	6.39	0.00	9.38	0.00
> 9	0.00	0.00	0.00	0.00

At each of the three time-points, the individual protocols of QLind- agents resulted in a higher percentage of patients with $[PO_4]_{serum}$ in the efficacy range and in a lower percentage of subjects in hyperphosphatemia than the clinical one. In particular, at the end of the fourth month, following the QL-based protocols 98% of patients was in the efficacy window (vs 70.21% for the clinical protocol). Prolonging the treatment, erdafitinib lost effectiveness for some patients in the partially responsive group, therefore a drop in the percentage of subjects having $[PO_4]_{serum} \in [5.5,7)\ mg/dL$ was observed for both the strategies (68.09% vs 56.73% for QL-based and clinical protocol, respectively). However, in the responsive patient subgroup, this decrease was not observed, confirming the outperformances of the QLind-agents in maintaining the $[PO_4]_{serum}$ in target.


A
Wilcoxon, p < 2.2e-16
% of Days with [PO4]serum < 5.5 mg/dL
100
90
80
70
60
50
40
30
20
10
0
Clinical Protocol
Q-Learning


B
Wilcoxon, p < 2.2e-16
% of Days with [PO4]serum in [5.5,7) mg/dL range
100
90
80
70
60
50
40
30
20
10
0
Clinical Protocol
Q-Learning

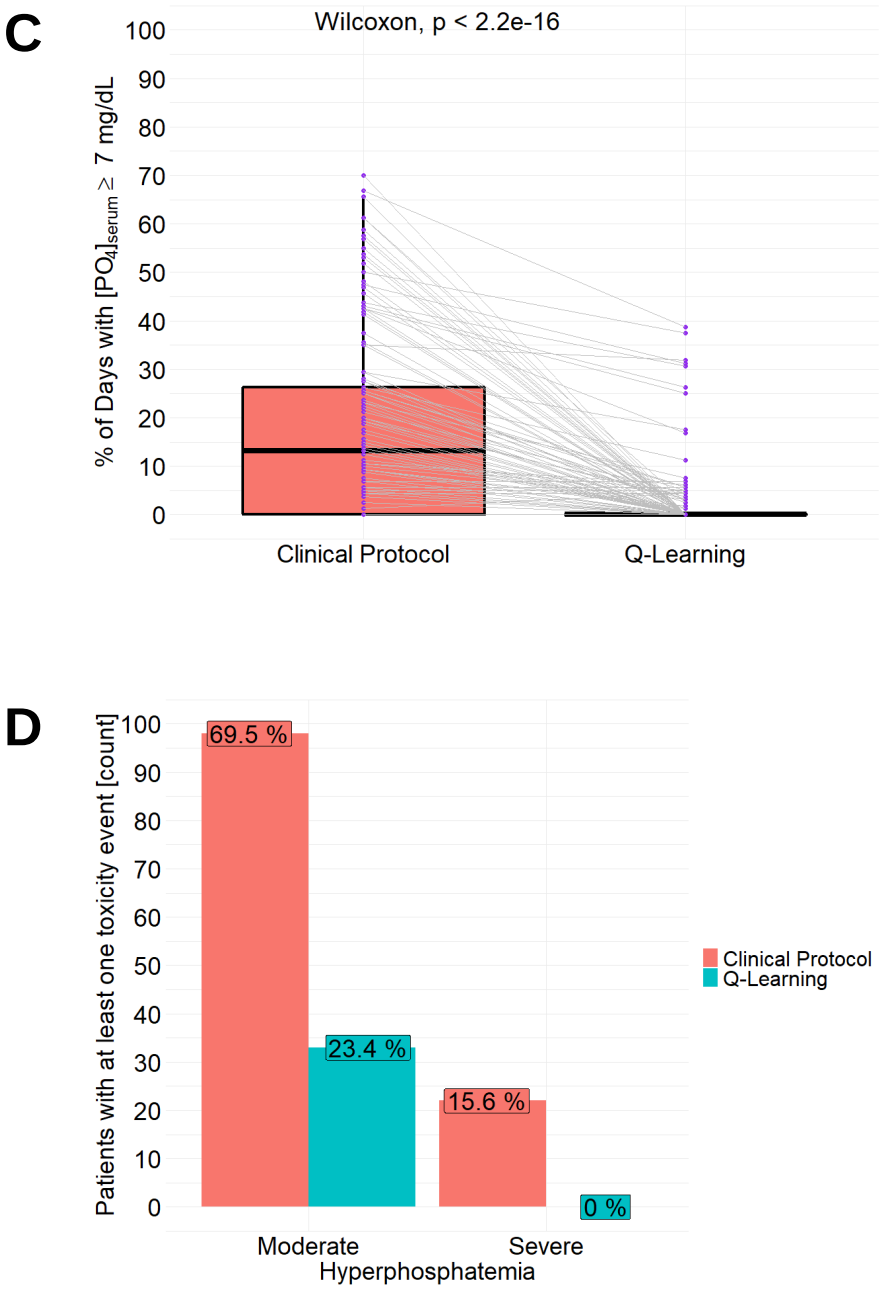


Figure 13: Summary of the performances of QL-based individual protocols. A-C) Paired comparisons of QLind-agents and the clinical protocol on the entire treatment time period for the percentages of days in which patients have $[PO_4]_{serum} < 5.5\, mg/dL$ (A), $[PO_4]_{serum} \in [5.5,7)\, mg/dL$ (B) and $[PO_4]_{serum} \geq 7\, mg/dL$ (Panel C) . For each of them, a Wilcoxon paired signed rank test ($\alpha = 0.05$) was performed. D) Paired comparisons of QLind-agents and the clinical protocol for the number of moderate and severe hyperphosphatemia events.

In addition, the efficacy and safety of the individual QL-based protocols and the clinical one were compared on the entire treatment period. In particular, paired analyses of the percentages of days in which individuals have $[PO_4]_{serum} < 5.5\ mg/dL$, $[PO_4]_{serum} \in [5.5,7)\ mg/dL$ and $[PO_4]_{serum} \geq 7\ mg/dL$ were performed at patient level applying a Wilcoxon paired signed rank test ($\alpha = 0.05$, H_0= median of the differences between percentages in the two groups is equal to 0). As illustrated in Figure 13 (Panels A-C), H_0 was always rejected. Therefore, on a treatment window of five months, QLind-agents were able to maintain $[PO_4]_{serum}$ in the target range for longer time, minimizing its permanence in both inefficiency and toxicity ranges. As concerns toxicity, the personal QL-based protocols completely avoided the onset of severe toxicity events and drastically reduced the number of moderate hyperphosphatemia occasions (Figure 13, Panel D). An example is reported in Figure 14.

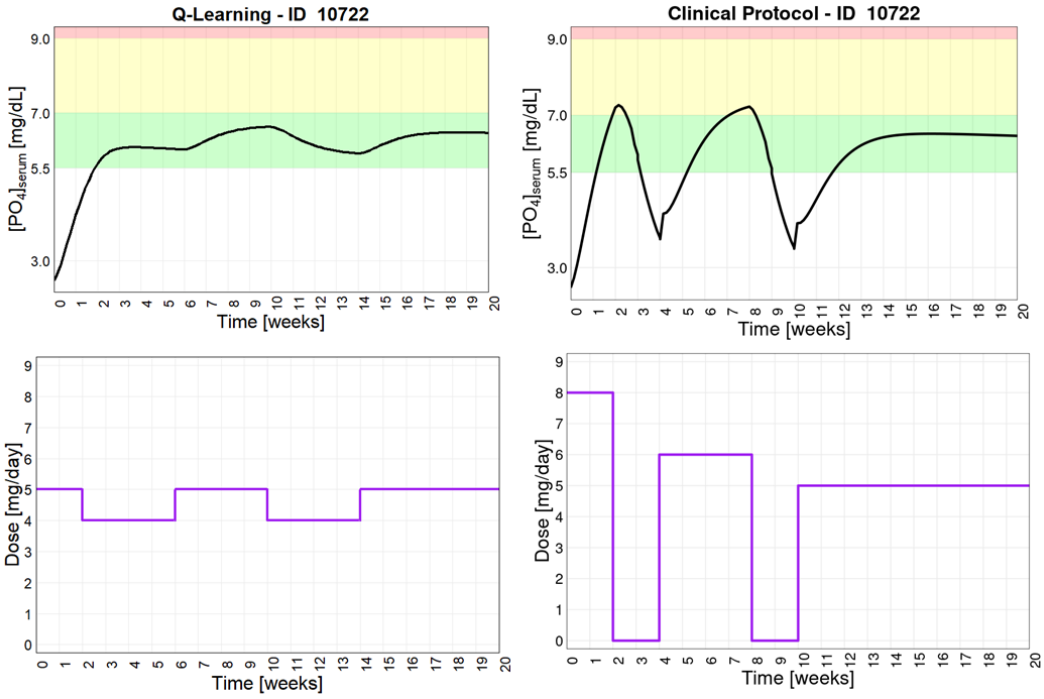


Figure 14: An example of how QL-based personalized protocol (left panels) brings to better $[PO_4]_{serum}$ levels avoiding the onset of moderate hypertension. This MIRL approach outperforms, in that patient, the FDA-approved clinical protocol (right panels).

The personalization of the initial dose significantly increased QLind-agents performances. As reported in Figure 15, the clinical initial dose of 8 mg/day was selected by the QLind-agents only in 23.40% of patients, while lower starting doses were generally preferred.

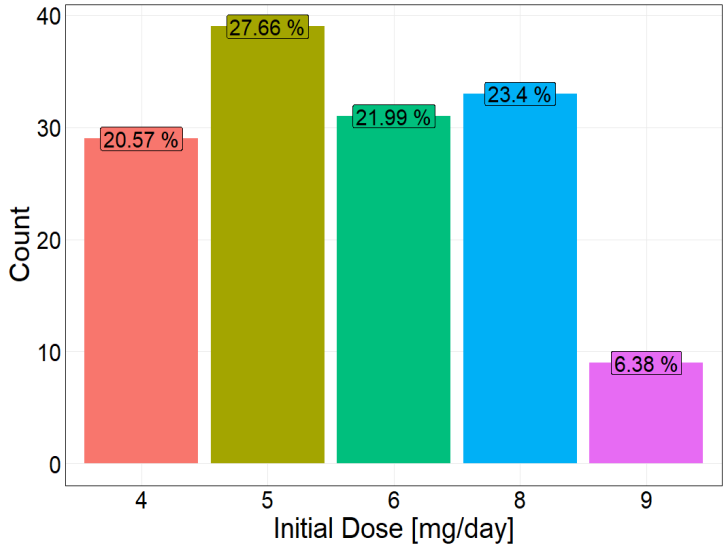


Figure 15: Distribution of erdafitinib starting doses selected by the individual QL-agents (8 mg/day is the clinical starting dose).

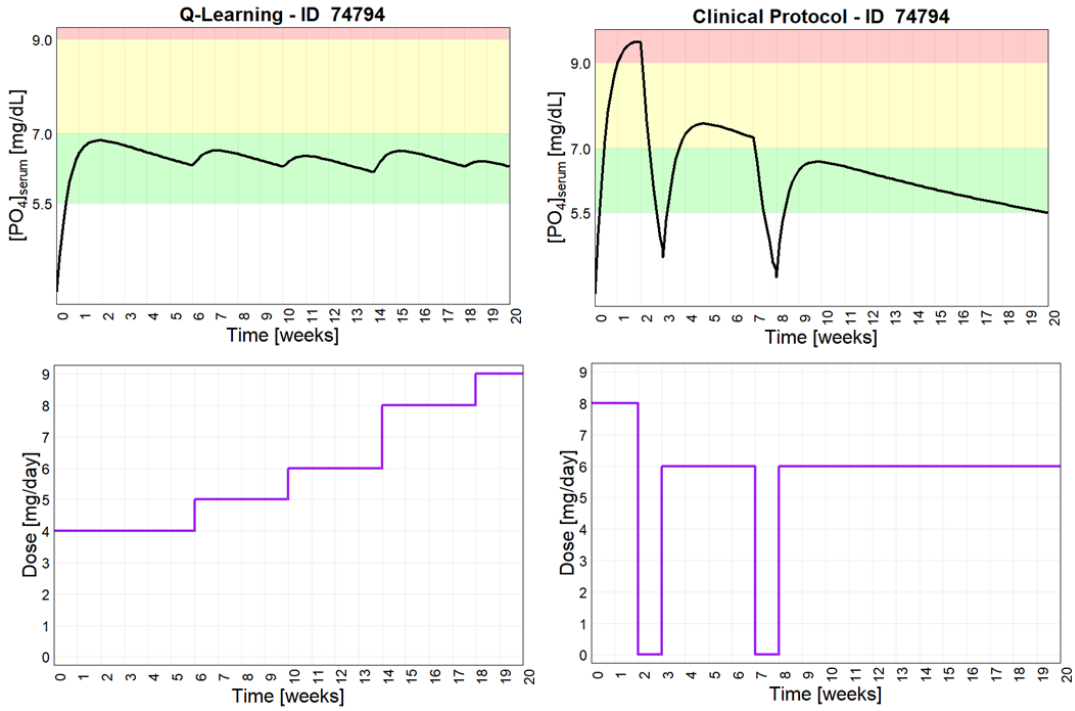


Figure 16: Comparison between erdafitinib dosing strategy of personal QL-agent (left panels) and that of FDA-approved clinical protocol (right panels). The QL-based dosing rules were able to optimize $[PO_4]_{serum}$ levels with a gradual up-titration of daily erdafitinib exposure. Differently, the clinical protocol led to severe hyperphosphatemia.

Figure 16 reported the example of a completely responsive patient with a 9 mg/day ideal dose: its optimized treatment started with a 4 mg/day dose (i.e., half of the dose specified by the protocol) that was then gradually increased up to 9 mg/day. This strategy allowed to circumvent the hyperphosphatemia events that occurred following the clinical protocol. A gradually dose increase was mainly adopted by QLind-agents also for partially responsive patients as it kept $[PO_4]_{serum}$ in target for longer time, until erdafitinib lost its effectiveness (Figure 17). These examples also underlined the strength of allowing QLind-agents to change the dose at each monitoring step.

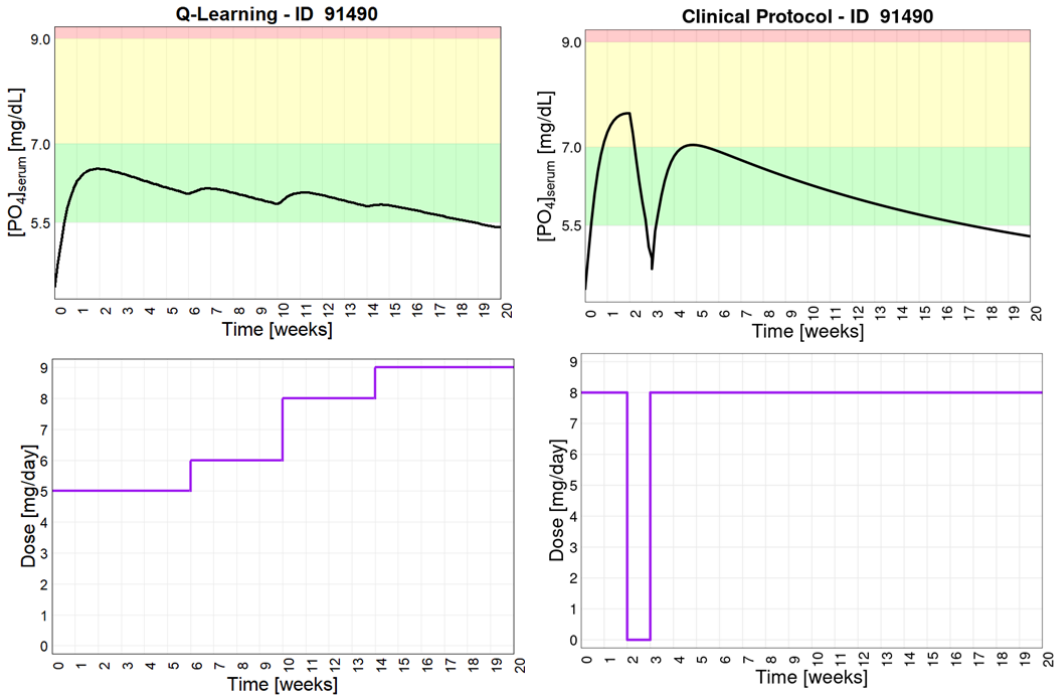


Figure 17: Comparison between the QL-based protocol and the clinical one for a partially responsive patient. QLind agent is able to keep $[PO_4]_{serum}$ within the normality range for longer time due to a gradual dose increase strategy.

Differently, for some patients, dose adjustments were not necessary, as the QLind-agents as they were able to detect the optimal dose from the beginning of the treatment. As illustrated in Figures 18 and 19, this behavior was found for patients treatable with either the highest or the lowest dose.

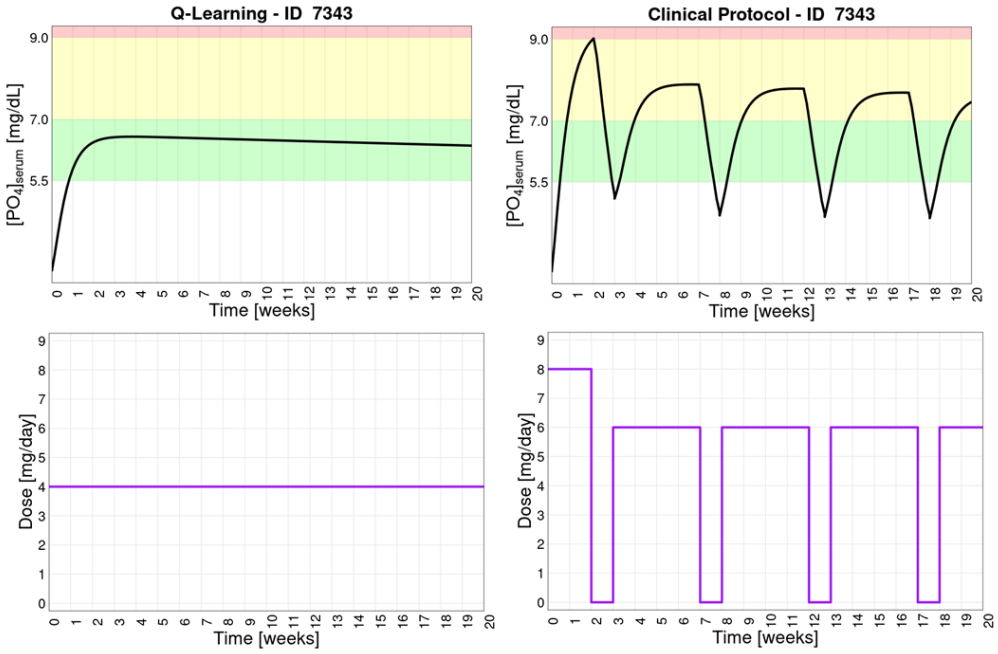


Figure 18: Example in which QLind-agent is able to discern that the patient can tolerate only low dosages. Therefore, $[PO_4]_{serum}$ is normalized by applying the lowest erdafitinib dosage (4 mg/day) since the beginning of treatment.

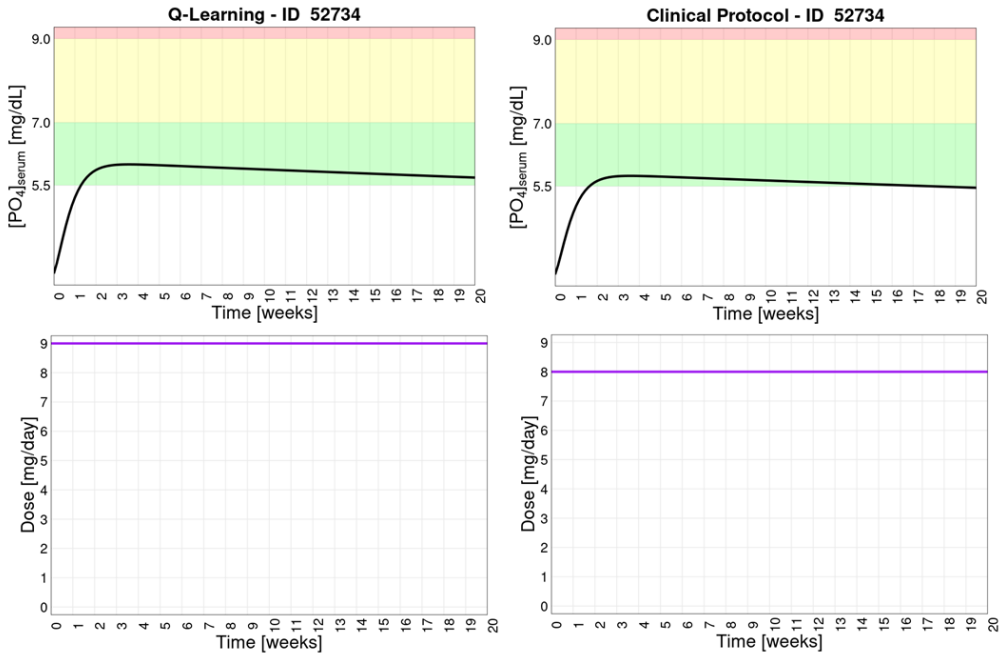


Figure 19: Example of how a QLind-agent is able to discern a patient for which $[PO_4]_{serum}$ is normalized only by applying the highest erdafitinib dosage (9 mg/day) from the beginning of treatment.

Interestingly, QLind-agents were also able to propose strategies based on a loading dose. For example, as shown in Figure 20, the personal QL-agent starts erdafitinib treatment with a dose of 8 mg/day. Then, starting from the second cycle, it is lowered to 6 mg/day. This preventive dose down-titration is more effective as it avoids moderate hyperphosphatemia. Differently, with the clinical protocol, hyperphosphatemia occurs as the dose of 8 mg/day is consecutively administered for the first three cycles.

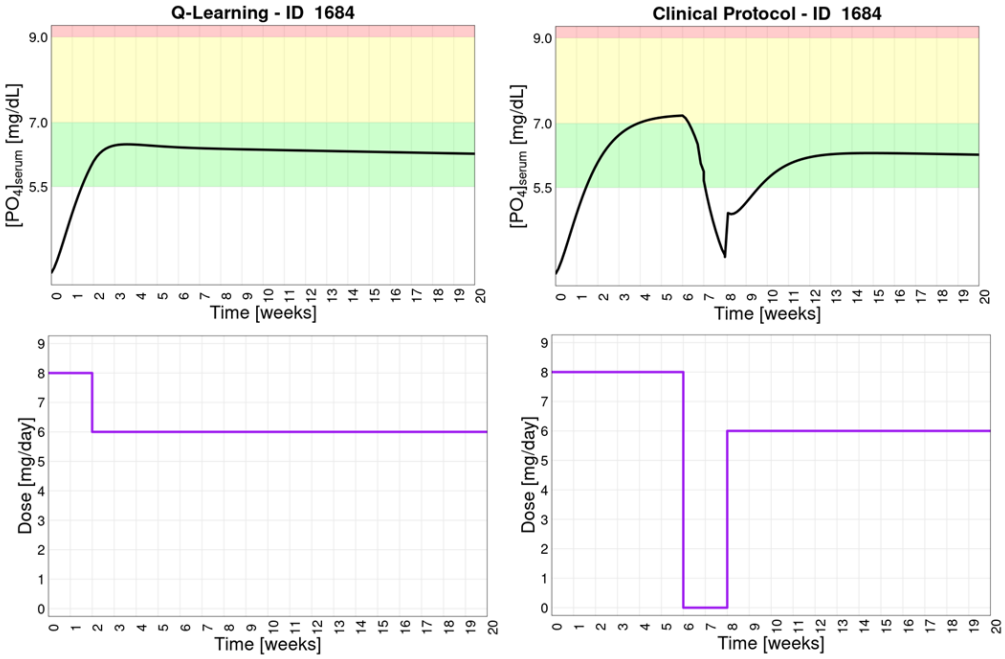


Figure 20: Example in which QL-agent selects an optimization strategy for erdafitinib based on a loading dose.

From the comparison between the adaptive dosing protocols defined by QLind-agents and the FDA-approved protocol, it emerged that the QL-based dosing strategies mainly managed moderate hyperphosphatemia adopting the same strategy of the clinical protocol. Indeed, in almost all the occasions of moderate hyperphosphatemia, QL-agents confirmed protocol decision of interrupting treatment rather than continuing with a lower or equal drug level (Figure 21).

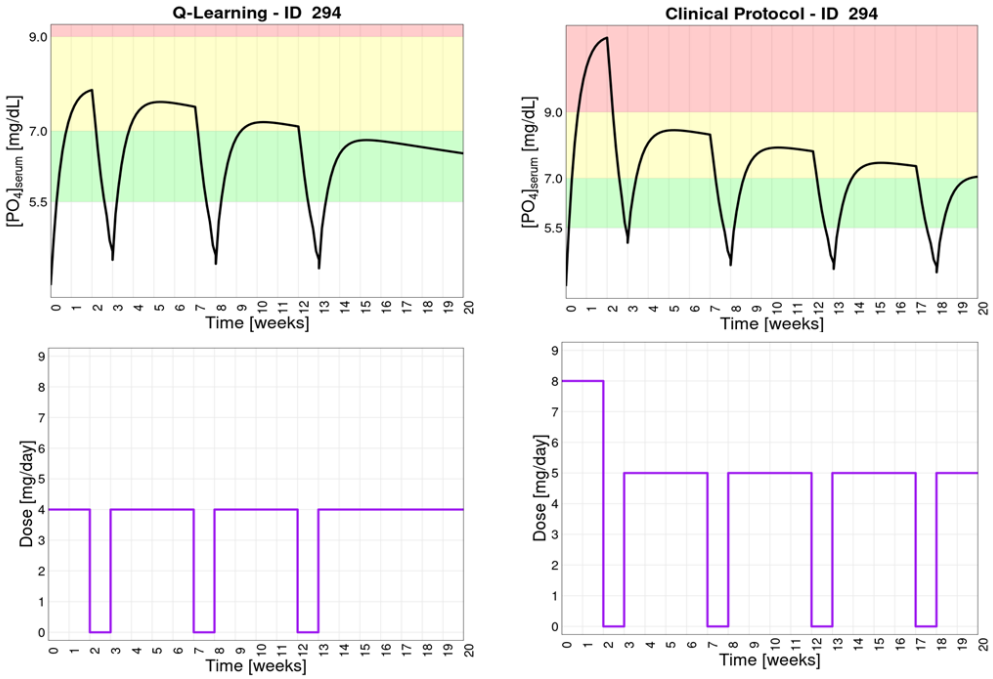


Figure 21.: Example in which QLind-agent decides to interrupt treatment in case of moderate hyperphosphatemia (yellow area) accordingly to the clinical protocol.

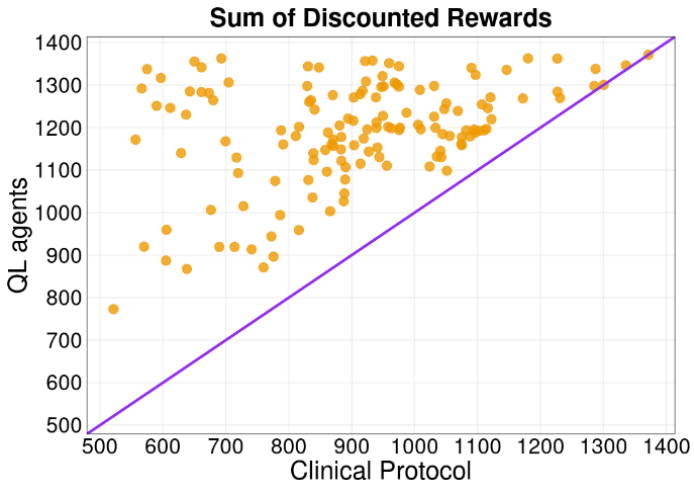


Figure 22: Comparison of the sum of discounted rewards between the customized dosing protocols administered by individual QL agent and the clinical dosing rules.

A final comparison between personal QL-based and FDA-approved protocols was made under the assumption that the reward function (Eqs.17-19) is a suitable measurement of the achievement of the therapeutic goal of erdafitinib therapy. Therefore, the sum of discounted rewards obtained with the QL-based individual and clinical protocols were compared and, as

reported in Figure 22, the QL-agents performed better, or at least equal, than the FDA-approved protocol.

3.2.2. RL general protocol vs individual RL-agents

To comprehensively evaluate the performances of the individual QL-based adaptive dosing protocols, a further benchmark comparison was conducted using the MIRL approach presented in section 2.2.1. To this end, a general erdafitinib adaptive dosing protocol optimized for a population of patients was derived by single QL-agent (i.e., QLpop-agent). The formalization of erdafitinib precision dosing problem presented in sections 3.1.2.-3.1.5. was maintained also for the setup of the QLpop-agent. Furthermore, also the previously introduced virtual population of 141 patients (details in section B.2. of Appendix B) was considered in this analysis. In particular, this cohort of virtual patients was intentionally used to train and test the QLpop-agent to allow the comparison between the two QL-based approaches and to put the QLpop-agent in the most favorable conditions.

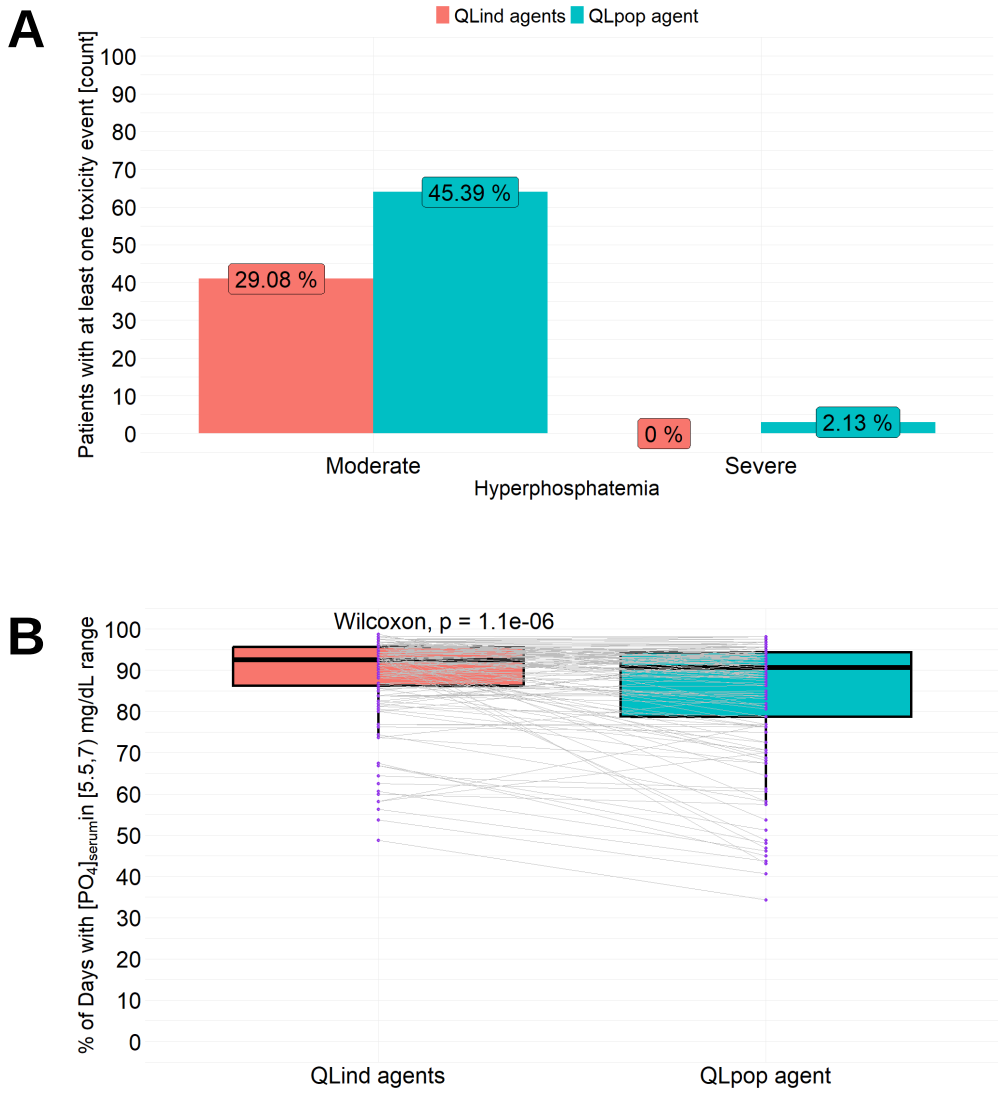
Indeed, by assuming that the training set coincides with the test set, the generalizability of the QLpop-protocol cannot be evaluated. Therefore, such approach would be incorrect for comparing the robustness with respect to inter-individual variability between the QLpop-rules and the erdafitinib clinical protocol. However, when the focus is on the comparison between the two QL approaches, evaluating the generalizability of QLpop in presence of protocols tailored on each patient by individual QL agents is not meaningful.

As showed in Table 3, at each endpoint, the QLpop-agent performed worser than the individual-based agents. This trend was consistent considering the percentages of patients in both the efficacy (i.e., $[PO_4]_{serum} \in [5.5,7)$ mg/dL) and toxicity (i.e., hyperphosphatemia defined as $[PO_4]_{serum} > 7$mg/dL) ranges. Furthermore, differently from individual QL-agents, the QLpop-agent was unable to completely avoid severe toxicities (i.e., $[PO_4]_{serum} > 9$ mg/dL) (Panel A of Figure 23). This clearly shows that the protocol learned by the QLpop-agent works well in average in the population but is not optimal for some of the patients.

Table 3: Comparison between QL-based individual protocols and those based on the QLpop-agent. At each clinically relevant time-point, the percentages of patients with $[PO_4]_{serum}$ within a certain range were used as evaluation metric. Results are reported for both the overall population and the subgroup of responsive patients.

Range of $[PO_4]_{serum}$ [mg /dL]	**%patients in the whole population [141 subjects]**		**%patients in the subgroup of completely responsive patients [96 subjects]**	
	QL-based individual protocols	**Population protocol of QLpop-agent**	**QL-based individual protocols**	**Population protocol of QLpop-agen**
Evaluation at the second week of treatment				
< 5.5	2.13	6.38	3.13	9.37
[5.5,7)	76.59	56.02	77.08	48.96
[7,9]	21.28	35.46	19.79	38.54
> 9	0.00	2.14	0.00	3.13
Evaluation at the end of the fourth month of treatment				
< 5.5	0.00	3.54	0.00	2.08
[5.5,7)	97.87	93.61	96.87	93.75
[7,9]	2.13	2.83	3.13	4.17
> 9	0.00	0	0.00	0.00
Evaluation at the end of treatment (fifth month)				
< 5.5	31.91	31.91	0.00	0.00
[5.5,7)	68.09	63.12	100	92.78
[7,9]	0.00	4.96	0.00	7.22
> 9	0.00	0.00	0.00	0.00

The gap between these two RL methodologies is further confirmed by looking at the entire treatment period. Indeed, the paired distributions of the percentages of days in which patients have $[PO_4]_{serum} \in [5.5,7)$ mg/dL and $[PO_4]_{serum} \geq 7$ mg/dL (Panels B and C of Figure 23) are in favour of the individual QL approach. Conversely, there is no statistical difference for what concerns the time in the inefficacy range (i.e., $[PO_4]_{serum} < 5.5$ mg/dL, Panel D of Figure 23).


A
QLind agents
QLpop agent
Patients with at least one toxicity event [count]
100
90
80
70
60
50
40
30
20
10
0
45.39 %
29.08 %
0 %
2.13 %
Moderate
Severe
Hyperphosphatemia
B
Wilcoxon, p = 1.1e-06
% of Days with [PO4]serum in [5.5,7) mg/dL range
100
90
80
70
60
50
40
30
20
10
0
QLind agents
QLpop agent

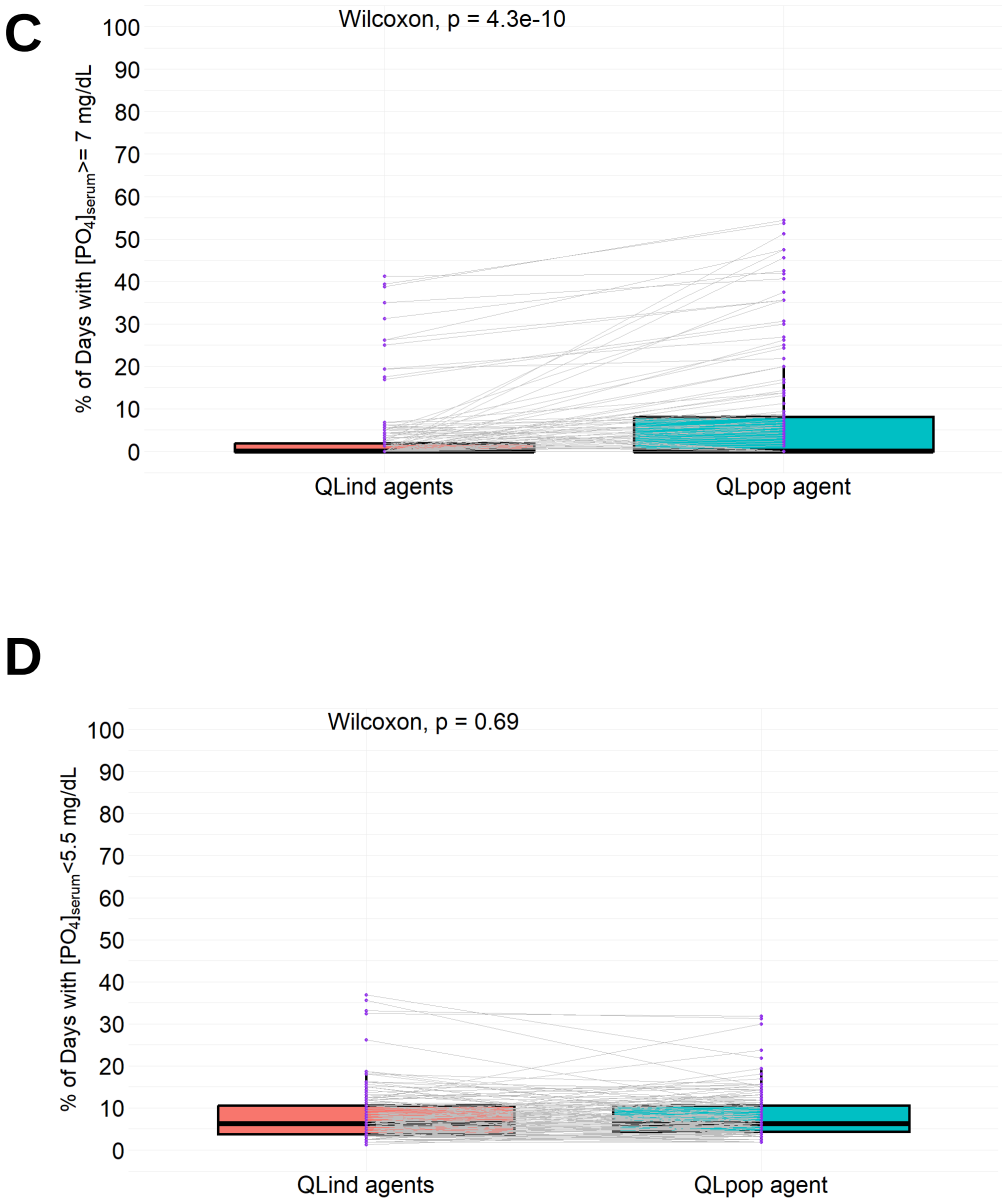


Figure 23: Summary of the comparison between individual QL-based protocols and those obtained by the QLpop-agent. A) Paired comparisons of individual QL-agents and the QLpop-agent for the number of moderate and severe hyperphosphatemia events. B-D) Paired comparisons of individual QL-agents and the QLpop-agent on the entire treatment time period for the percentages of days in which patients have $[PO_4]_{serum} \in [5.5,7)\ mg/dL$ (B), $[PO_4]_{serum} \geq 7\ mg/dL$ (C) and $[PO_4]_{serum} < 5.5\ mg/dL$ (D).

3.3. Discussions

In this Chapter, a first application of the novel MIRL paradigm (section 2.2.2.) to tailor adaptive dosing strategies on each individual was presented. In particular, this framework was applied on a relevant case study directly derived from clinical oncology, i.e., the erdafitinib treatment of metastatic urothelial carcinoma. An adaptive dosing protocol (Figure 10) in which erdafitinib dose is adjusted based on $[PO_4]_{serum}$, the efficacy/safety biomarker of this treatment, has already been approved by FDA. However, due to the narrow therapeutic window and significant IIV of pharmacological response [100], erdafitinib therapy could potentially benefit from a personalization of the adaptive dosing protocol at individual patient level. To this end, the novel MIRL approach was applied using QL as RL algorithm, and the obtained results confirmed the potentiality of this methodology in supporting treatment personalization. This successful result relies on some fundamental steps that are here discussed.

First, the clinical precision dosing problem must be correctly formalized in terms of elements composing the RL-framework, that for QL are a discretized representation of system states, agent actions and reward function. More in detail, system states must include all the information relevant to take the right action to perform (dose selection in this study). For the erdafitinib case study, states included the discretized $[\mathrm{PO_4}]_{\mathrm{serum}}$ level representing the patient status but also stored information on the previous administered dose and the hyperphosphatemia level that caused temporary treatment interruption. In particular, $[\mathrm{PO_4}]_{\mathrm{serum}}$ was discretized in six levels, one for inefficacy, three for efficacy, two for moderate and severe toxicity. Splitting the efficacy $[\mathrm{PO_4}]_{\mathrm{serum}}$ range in multiple sublevels allowed the QL-agent to select a different action depending on whether the biomarker lies in the bottom, middle or top of the target window. This was fundamental to adequately manage the loss of erdafitinib efficacy over time. Regarding the definition of agent actions, they must account for relevant clinical constraints, for example due to safety reasons, in order to obtain QL-based protocols acceptable in the clinical practice

At the same time, agents must have a certain degree of freedom in taking decisions to allow a treatment customization on individual patients. For erdafitinib, according to the clinical protocol, QL-agents were imposed to select only clinically available doses, to gradually increase/decrease the dose and to temporally interrupt the treatment in case of severe hyperphosphatemia. Instead, differently from the clinical protocol, QL agents were allowed to select a personalized initial dose potentially different from $8\,mg/day$, to gradually adjust erdafitinib dose at each monitoring occasion, to decrease the dose without interrupting the treatment in presence of moderate toxicities, and to resume treatment after interruption if the hyperphosphatemia severity is decreased at least by one level. The set of actions that the agent can perform was designed with the aim of obtaining

clinically acceptable protocols. However, their choice was arbitrary, and it could be of interest to train QL-agents with a different degree of freedom.

Finally, the reward function (Eqs. 17-19) must suitably encode the therapeutic scope of the treatment to correctly score the different actions and optimize the adaptive dosing protocol. For erdafitinib, the goal was to maintain $[PO_4]_{serum}$ within its efficacy range (i.e., $[5.5,7)$ mg/dL) as long as possible, avoiding hyperphosphatemia events. The reward function was designed to faithfully summarize the expert knowledge on erdafitinib treatment available from the literature [99–101,104,107]. Its form was selected based on other case studies that will be discussed in the next chapters and it does not simply describe the difference from a target value [83,86,88]. As the adopted reward function led to satisfactory results, a fine-tuning process was not performed. However, in view of an actual clinical application, the design of the reward function, as well as the entire formalization of the proposed RL-framework, should be carefully reviewed with clinicians to ensure that all the key aspects of the erdafitinib treatment are correctly considered and formalized. Furthermore, as will be discussed in the next chapters, in presence of a richer modelling framework the reward could be further expanded. For example, terms linked to the probability of survival or of adverse reactions can be included, thus leading to a multi-objective treatment optimization (e.g., joint model of multiple treatment biomarkers and overall survival [108]).

Once the RL-based precision dosing framework has been defined, it must be tested on a heterogeneous virtual population, in terms of pharmacological response. Indeed, evaluating the RL performances on very different scenarios is fundamental to assess its flexibility with respect to the IIV, a central feature in precision dosing whose goal is finding the right dosage rules for every patient. In this explorative analysis, a heterogenous virtual population (n=141) was *ad hoc* constructed equally sampling patients from both completely (n=96) and partially (n=45) responsive groups further stratified by their ideal dose levels (further details on its generation are reported in section B.2. of Appendix B). This strategy allowed to design a virtual population including a wide range of different patients, maintaining a limited sample size.

When the novel MIRL approach was applied on each individual of the virtual population, the individual QL-based protocols outperformed the FDA-approved one in terms of both efficacy and safety. They were able to lead $[PO_4]_{serum}$ in the target range in a higher patient percentage and to maintain the biomarker in target for a longer time-period respect to the clinical protocol. In addition, the severe hyperphosphatemia events were completely avoided, and the moderate ones drastically reduced (Figure 13 and Table 2). The personalization of the starting dose was fundamental for the nice performances of the individual QL-agents (QLind-agents) that were able to detect patients tolerating the highest dose as well as those requiring a low initial drug amount or a loading dose (Figures 15 and 20). These findings suggested the individualization of erdafitinib starting dose could

significantly contribute to reduce toxicity events, especially at the beginning of the treatment (Figures 17 and 18). Further, also a gradual dose variation at each monitoring step, even when the biomarker falls in the target window, was a further key feature of QLind-agents that allowed a better and more far-sighted control of $[PO_4]_{serum}$. This was a deviation from the clinical strategy which is based on pushing high doses and then lowering them if toxicity events occur (Figures 14, 16 and 17). Differently, in almost moderate hyperphosphatemia events, QLind-agents followed the protocol decision of interrupting treatment rather than continuing with a lower or equal dose (Figure 21).

Overall, the obtained finding demonstrated that the erdafitinib treatment could significantly benefit from a precision dosing strategy in which dose adjustment rules are individually tailored on each individual patient. This conclusion was further confirmed by comparing the individual QL-based protocols here identified by training a QLind-agent for each patient with an optimal QL-based adaptive protocol obtained by training a single QL-agent on the entire patient population (QLpop-agent). The protocol learned by the QLpop-agent performed worser than the QLind-agents (Figure 23 and Table 3). Indeed, the output of the QLpop-agent is a general dosing strategy that is not individually tailored on each subject but provides individualized dosing indications learned by evaluating their effect on a group of patients rather than on a single one. Therefore, it can work well for the majority of the population but cannot be the optimal solution for some of the patients. This limitation is not imputable to the RL approach itself, but it is intrinsic to every dosing strategy designed for a whole population.

Despite the promising results on erdafitinib case study, this novel patient-centric MIRL method relies on some relevant assumptions that partially hinder its immediate clinical applicability. Therefore, these aspects that have to be accurately addressed before the presented findings could successfully translate into an actual clinical application.

First of all, it was hypothesized that the PK-PD model perfectly described the erdafitinib pharmacological response and RUV was not considered in the simulations. Neglection of RUV is a strong simplification that was necessary for a preliminary comprehensible evaluation of the proposed novel MIRL methodology. Indeed, it is known that effectiveness of RL algorithms potentially decreases when RUV is introduced [73]. Consequently, RUV must be adequately managed, with solutions that must depend on what RUV represents, e.g., measurement errors or model misspecification. In the first case, moving from MDP to Partially Observable MDP [109] could be a possible strategy to formally describe the uncertainty of system state, due to RUV. Model misspecification should be addressed differently, as will be discussed in Chapter 6. In this case, the reward function would be no longer deterministic, and it could be necessary to switch from classical RL algorithms (e.g., QL) to different version that properly handles this stochastic framework [76,93].

Furthermore, it was assumed that digital patient twins were well-characterized since the beginning of the treatment (i.e., the individual PK-PD model parameters were known). However, individual parameter estimation requires the availability of individual data and, consequently, needs to be performed during the monitoring of the treatment. As will be discussed in Chapter 6, the integration of the QL-based treatment optimization within a Bayesian paradigm allowing a continuous learning of model parameters from the monitoring observations [63,80,98], could definitively cope with this issue. In this scenario, at each monitoring step, the individual model used as specific patient digital twin, and the personal RL-agent should be updated accounting for new patient data. This means that the personal RL-agent has to be continuously re-trained accordingly on the new available patient knowledge. As this process makes a massive usage of model simulations, it can be very time-consuming in presence of a complex modelling framework. The duration of the retraining process, depending on the clinical workflow, has to fall within the time frame spanning from patient observation to clinical decision in order to leverage RL-based suggestion. Consequently, an efficient implementation of the algorithm and the modelling framework on a powerful computing architecture are needed.

In conclusion, the novel MIRL approach to support patient-tailored adaptive dosing strategies showed promising results. Simulations demonstrated that the RL-based personalized strategies were superior to the FDA-approved dosing protocol. The next chapters will cover additional applications of this methodology in more complex precision dosing scenarios, as well as new strategies to overcome the limitations highlighted in this section.

Chapter 4

Joint optimization of multiple treatment biomarkers. Application of the RL/PK-PD framework to givinostat therapy in polycythemia vera patients [3]

Most of the literature applications of MIRL for precision dosing are focused on deriving a general administration protocol for an entire population. In particular, the majority of these studies consider only a single biomarker of treatment efficacy and/or toxicity [39,73,75,83,85–89,92]. However, real clinical settings are generally more complicated due to multiple efficacy/toxicity biomarkers to simultaneously optimize during pharmacotherapy [39,74]. For example, although tumor shrinkage is the primary goal of chemotherapy, also the alteration of haematopoietic and renal/hepatic functions are monitored and considered for dose adjustments to avoid abnormal values resulting harmful for patients.

The aim of this chapter is to apply MIRL to multi-objective precision dosing, with the dual goal of deriving a clinically acceptable general adaptive dosing protocol for a population of patients and individually tailored strategies. Furthermore, here for the first time, the potentialities of RL-based dose personalization are investigated to optimize the drug development process. Finally, this study explores a novel hybrid MIRL implementation framework, offering a methodological alternative to those discussed in section 3.1.2.4.

[3] The content of this chapter is under review for a publication on Clinical Pharmacology and Therapeutics.

To this end, the multi-objective precision dosing problem of givinostat (Italfarmaco, ITF2357), a compound under clinical development for the treatment of polycythemia vera (PV), was considered as case study [110]. Givinostat treatment inhibits the uncontrolled myeloproliferation of bone marrow cells in PV patients, and the monitoring of platelets (PLT), neutrophils (or white blood cells, WBC) and haematocrit (HCT) is used to assess both therapy efficacy and toxicity as well as to guide dose adjustments.

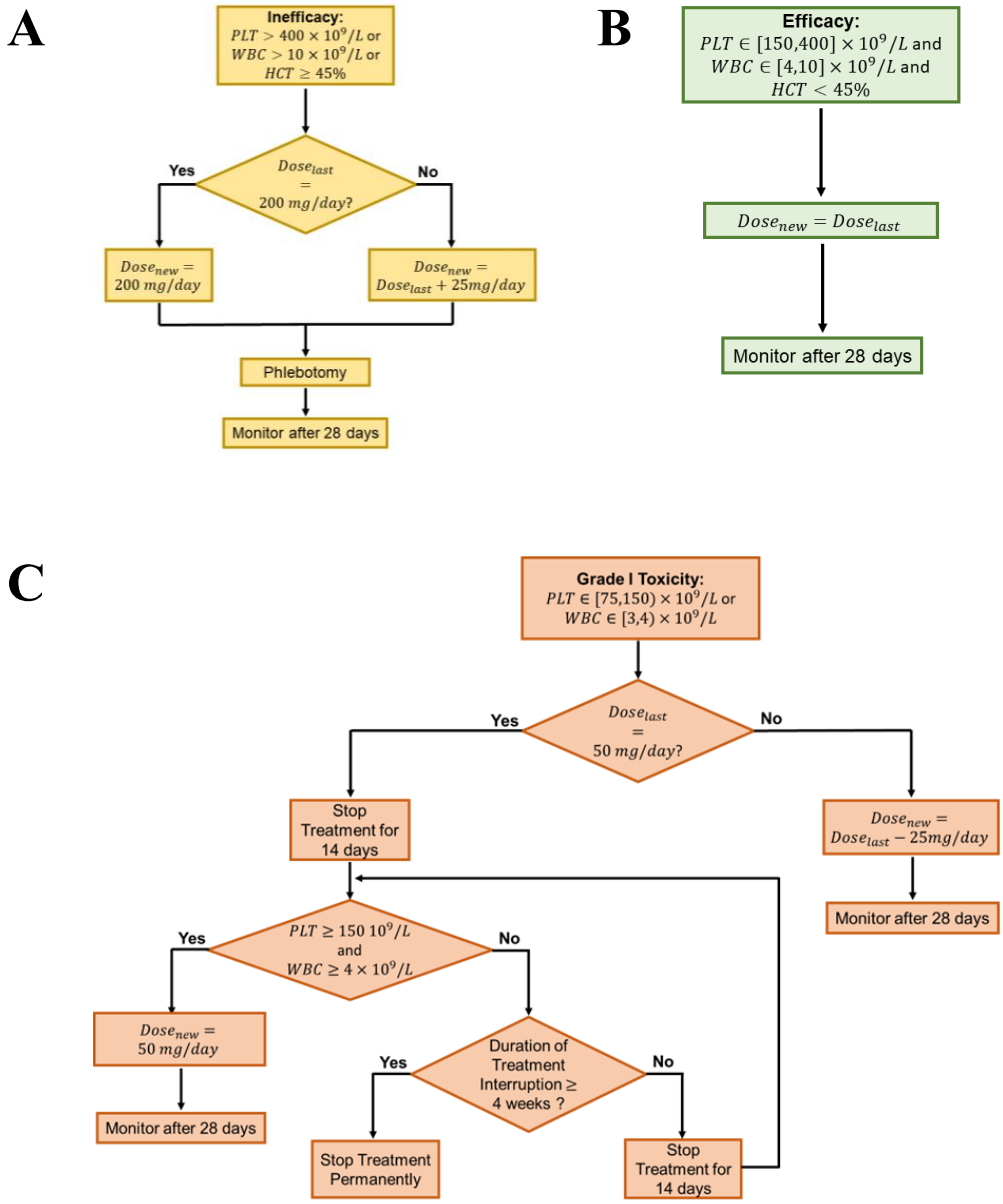
This chapter is structured as follows. First, givinostat and its adaptive dosing protocol will be presented in section 4.1.1. Then, the attention will be shifted to translating givinostat clinical setup within RL, and in particular QL. Therefore, section 4.1.2 will provide a description of the implemented QL setup and its evaluation framework. Finally, sections 4.2 and 4.3 will present and discuss the obtained results. Supplementary information of this chapter is reported in Appendix C.

4.1. Methods

4.1.1. Givinostat treatment of polycythemia vera

Givinostat is a drug under clinical evaluation for the treatment of PV [110], a chronic life-threatening neoplasm characterized by an overproduction of blood cells which increases the levels PLT, WBC and HCT [111–113]. The goal of givinostat treatment is to contrast the uncontrolled myeloproliferation and, consequently, to simultaneously induce and maintain the three haematological parameters within an acceptable range (i.e., $PLT \in [150{,}400] \times 10^9/L$, $WBC \in [4{,}10] \times 10^9/L$, $HCT < 45\%$) [45]. PLT, WBC and HCT are key endpoints for both efficacy and safety of givinostat treatment. Indeed, the treatment results ineffective if the haematological parameters are above the upper limits, instead, it induces some toxicity (i.e., thrombocytopenia and/or neutropenia) if they decrease below the lower limits. Toxicities can be of Grade 1 (i.e., moderate) if $PLT \in [75{,}150) \times 10^9/L$ and/or $WBC \in [3{,}4) \times 10^9$/L, or Grade 2 (i.e., severe) if $PLT < 75 \times 10^9/L$ and/or $WBC < 3 \times 10^9/L$. To maintain the hematological parameters within the efficacy range, a dose-adaptive administration protocol based on the monitoring of PLT, WBC and HCT was proposed for the planned Phase III trial on givinostat [39]. Givinostat can be administered at {50,75,100,125,150,175,200} mg/day dose level. For each patient, the treatment starts with a fixed dose of 100 mg/day. Only patients with a baseline $HCT \geq 45\%$ receive a preliminary phlebotomy to normalize this haematological parameter. Then, at the end of each 28-day cycle the dose is adjusted (i.e., confirmed or increased/decreased by 25 mg/day) based on the measured PLT, WBC, and HCT values following rules schematized in

Figure 24. In particular, the treatment is interrupted for 14 days (Figure 24, panels C-D) if severe toxicities occur.



D

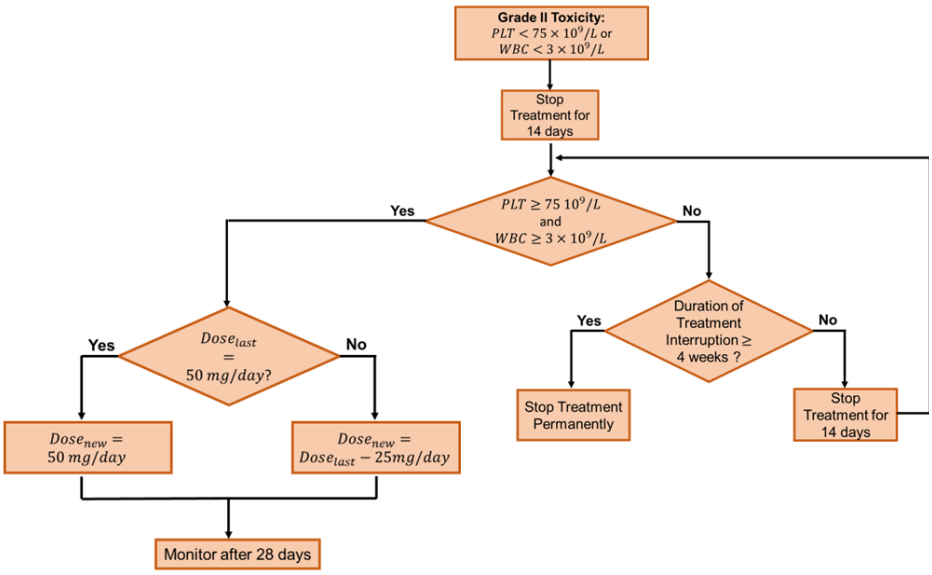


Figure 24: Givinostat adaptive-dose protocol flow charts. Panel A illustrates the rules in case of inefficacy, Panel B in presence of efficacy. Panels C and D report the decisional steps made in presence of Grade I and Grade II toxicity, respectively.

4.1.2. Setup of QL algorithm for givinostat precision dosing

An essential preliminary step to apply the MIRL framework is to formalize givinostat precision dosing problem as a MDP (section 2.1.1). Therefore, the key elements of an MDP, that is, system states, agent actions and reward function, were defined in accordance with givinostat clinical setting described in section 4.1.1. Since QL was used as RL algorithm to optimize givinostat adaptive dosing protocol at both single patient and population levels, the finite MDP framework (section 2.1.1) was adopted.

In particular, system/patient state are based on the PLT, WBC and HCT levels which are periodically monitored to guide dose adjustment (i.e., QL-agents action). As the clinical goal is to achieve and maintain a Complete Haematological Response (CHR), which corresponds to simultaneously have $PLT \in [150{,}400] \times 10^9/L$, $WBC \in [4{,}10] \times 10^9/L$ and $HCT < 45\%$ [45], dose adjustments are evaluated by a reward function accounting for all the three hematological parameters. QL algorithm was integrated with either a virtual patient or a population of patients, depending on whether the task was to derive individually tailored adaptive dosing protocols or a general protocol.

The following subsections will provide a comprehensive description of the QL setup adopted for givinostat precision dosing problem and its implementation.

4.1.2.1. Reward function

Treatment goal is to induce and maintain the three haematological parameters within their target ranges, in particular avoiding severe toxicities (i.e., $PLT < 75 \times 10^9/L$ and/or $WBC < 3 \times 10^9/L$). Therefore, in presence of severe myelosuppression, the reward function returns 0 (i.e., the lowest reward value), otherwise it is given by the sum of three terms ($Reward_{PLT}$, $Reward_{WBC}, Reward_{HCT}$), one for each hematological parameter.

$$Reward = \begin{cases} 0 & \text{if } PLT_{Obs} < 75 \times 10^9/L \text{ and/or } WBC_{Obs} < 3 \times 10^9/L \\ Reward_{PLT} + Reward_{WBC} + Reward_{HCT} & \text{otherwise.} \end{cases}$$

(23)

Each of the three terms in Eq. 23 is given, in turn, by a weighted sum of three functions that take values in the range [0,1]:

$$\begin{cases} Reward_{PLT} = \beta_1 \cdot Reward_{PLT,Obs} + \beta_2 \cdot Reward_{PLT,Range} + \beta_3 \cdot Reward_{PLT,Der} \\ Reward_{WBC} = \beta_1 \cdot Reward_{WBC,Obs} + \beta_2 \cdot Reward_{WBC,Range} + \beta_3 \cdot Reward_{WBC,Der} \\ Reward_{HCT} = \beta_1 \cdot Reward_{HCT,Obs} + \beta_2 \cdot Reward_{HCT,Range} + \beta_3 \cdot Reward_{HCT,Der} \\ \text{with } \beta_1 = 10, \ \ \beta_2 = 7 \text{ and } \beta_3 = 5. \end{cases}$$

(24)

The weighting strategy is the same for the three haematological parameters. The highest importance is given to the term $Reward_{Obs}$ (Eqs.25-27) which scores the haematological parameter level (i.e., PLT_{Obs}, WBC_{Obs}, HCT_{Obs}) at the end of each treatment cycle with respect to the therapeutic window. Actions bringing the haematological parameter as close as possible to the middle of the normal range are more remunerated.

$$Reward_{PLT,Obs} = \begin{cases} a \cdot e^{-0.01 \cdot |PLT_{obs} - 150|} + c & \text{if } PLT_{Obs} \in [75,150) \times 10^9/L \\ 0.5 \cdot (1 - e^{-0.06 \cdot (PLT_{Obs} - 150)} + 0.5 & \text{if } PLT_{Obs} \in [150,275) \times 10^9/L \\ 0.5 \cdot \left(1 - e^{-0.06 \cdot |PLT_{Obs} - 150|}\right) + 0.5 & \text{if } PLT_{Obs} \in [275,400) \times 10^9/L \\ 0.5 \cdot e^{-0.01 \cdot (PLT_{Obs} - 400)} & \text{if } PLT_{Obs} \geq 400 \times 10^9/L \\ a = 0.643 & c = -0.143 \end{cases}$$

(25)

$$Reward_{WBC,Obs} = \begin{cases} a \cdot e^{-0.01 \cdot |37.5 \cdot WBC_{Obs} - 150|} + c & \text{if } WBC_{Obs} \in [3,4) \times 10^9/L \\ 0.5 \cdot \left(1 - e^{-0.06 \cdot (37.5 \cdot WBC_{Obs} - 150)}\right) + 0.5 & \text{if } WBC_{Obs} \in [4,7) \times 10^9/L \\ 0.5 \cdot \left(1 - e^{-0.06 \cdot |40 \cdot WBC_{Obs} - 150|}\right) + 0.5 & \text{if } WBC_{Obs} \in [7,10) \times 10^9/L \\ 0.5 \cdot e^{-0.01 \cdot (40 \cdot WBC_{Obs} - 400)} & \text{if } WBC_{Obs} \geq 10 \times 10^9/L \\ a = 0.643 & c = -0.143 \end{cases}$$

(26)

$$Reward_{HCT,Obs} = \begin{cases} 0.5 \cdot \left(1 - e^{-0.06 \cdot (8.89 \cdot HCT_{Obs} - 150)}\right) + 0.5 & \text{if } HCT_{Obs} < 22.5\% \\ 0.5 \cdot \left(1 - e^{-0.06 \cdot |8.89 \cdot HCT_{Obs} - 400|}\right) + 0.5 & \text{if } HCT_{Obs} \in [22.5,45)\% \\ 0.5 \cdot e^{-0.01 \cdot (8.89 \cdot HCT_{Obs} - 400)} & \text{if } HCT_{Obs} \geq 45\% \end{cases}$$

(27)

The second term, $Reward_{Range}$, weighted by β_2, considers the proportion of days in a treatment cycle in which the haematological parameter is within its target range (Eqs. 28-30). This contribution captures the need of normalizing the haematological parameter as long as possible and not only at the end of the treatment cycle.

$$Reward_{PLT,Range} = \begin{cases} \dfrac{\sum_i PLT_{cycle,i}}{length(cycle)} \\ PLT_{cycle,i} = 1 \quad \text{if } PLT_{cycle,i} \in [150,400] \times 10^9/L \\ PLT_{cycle,i} = 0 \quad \text{if } PLT_{cycle,i} \notin [150,400] \times 10^9/L \end{cases}$$

(28)

$$Reward_{WBC,Range} = \begin{cases} \dfrac{\sum_i WBC_{cycle,i}}{length(cycle)} \\ WBC_{cycle,i} = 1 \quad \text{if } WBC_{cycle,i} \in [4,10] \times 10^9/L \\ WBC_{cycle,i} = 0 \quad \text{if } WBC_{cycle,i} \notin [4,10] \times 10^9/L \end{cases}$$

(29)

$$Reward_{HCT,Range} = \begin{cases} \dfrac{\sum_i HCT_{cycle,i}}{length(cycle)} \\ HCT_{cycle,i} = 1 \quad \text{if } HCT_{cycle,i} < 45\% \\ HCT_{cycle,i} = 0 \quad \text{if } HCT_{cycle,i} \geq 45\% \end{cases}$$

(30)

The last term, $Reward_{Der}$, weighted by β_3, is a function of the derivative, y', of the corresponding haematological parameters, where y' is approximated trough difference between the two last values of a treatment cycle. Within the target range, this function (Eqs. 31-33 and Figures C.1-3 of Appendix C) gives higher remuneration to stable trend (i.e., y' close to 0). Conversely, in presence of toxicity or inefficacy, it returns a higher reward if y' is positive or negative, respectively. In this way, also actions driving the haematological parameters to their target range are fostered.

$$Reward_{PLT,Der}(y'_{PLT}) = \begin{cases} \dfrac{0.2}{1+e^{-y'}} & \text{if } y' \leq 0 \text{ and } PLT_{obs} < 150 \times 10^9/L \\ \dfrac{1}{1+e^{-y'}} & \text{if } y' > 0 \text{ and } PLT_{Obs} < 150 \times 10^9/L \\ e^{\frac{y^2}{\psi}}, \quad \psi = 1/\ln(5) & \text{if } PLT_{Obs} \in [150,400] \times 10^9/L \\ \dfrac{1}{1+e^{-|y'|}} & \text{if } y' < 0 \text{ and } PLT_{Obs} > 400 \times 10^9/L \\ \dfrac{0.2}{1+e^{-y'}} & \text{if } y' \geq 0 \text{ and } PLT_{Obs} > 400 \times 10^9/L \end{cases}$$

(31)

$$Reward_{WBC,Der}(y'_{WBC}) = \begin{cases} \frac{0.2}{1+e^{-y'}} & \text{if } y' \leq 0 \text{ and } WBC_{Obs} < 4 \times 10^9/L \\ \frac{1}{1+e^{-y'}} & \text{if } y' > 0 \text{ and } WBC_{Obs} < 4 \times 10^9/L \\ e^{\frac{y^2}{\psi}}, \quad \psi = 1/\ln(5) & \text{if } WBC_{Obs} \in [4,10] \times 10^9/L \\ \frac{1}{1+e^{-|y'|}} & \text{if } y' < 0 \text{ and } WBC_{Obs} > 10 \times 10^9/L \\ \frac{0.2}{1+e^{-y'}} & \text{if } y' \geq 0 \text{ and } WBC_{Obs} > 10 \times 10^9/L \end{cases}$$

(32)

$$Reward_{HCT,Der}(y'_{HCT}) = \begin{cases} e^{\frac{y^2}{\psi}}, \quad \psi = 1/\ln(5) & \text{if } HCT_{Obs} < 45\% \\ \frac{1}{1+e^{-|y'|}} & \text{if } y' < 0 \text{ and } HCT_{Obs} \geq 45\% \\ \frac{0.2}{1+e^{-y'}} & \text{if } y' \geq 0 \text{ and } HCT_{Obs} \geq 45\% \end{cases}$$

(33)

4.1.2.2. System/Patient states

The system (i.e., patient) state is described by a tuple of four elements, $X = \{PLT_{Discr}, WBC_{Discr}, HCT_{Discr}, PrevDose\}$.

$PLT_{Discr.}, WBC_{Discr.}, HCT_{Discr.}$ (Eqs. 34-36) are the observations of PLT, WBC and HCT at the monitored day ($PLT_{Obs}, WBC_{Obs}, HCT_{Obs}$) categorized according to the inefficacy, efficacy and toxicity definition suggested by the clinico-hematological European Leukemia Network criteria [114].

$$PLT_{Discr}(PLT_{Obs}) = \begin{cases} 1 & \text{if } PLT_{Obs} < 75 \times 10^9/L \ (Severe\ Thrombocytopenia) \\ 2 & \text{if } PLT_{Obs} \in [75,150) \times 10^9/L \ (Moderate\ Thrombocytopenia) \\ 3 & \text{if } PLT_{Obs} \in [150,400] \times 10^9/L \ (Efficacy) \\ 4 & \text{if } PLT_{Obs} > 400 \times 10^9/L \ (Unefficacy) \end{cases}$$

(34)

$$WBC_{Discr}(WBC_{Obs}) = \begin{cases} 1 & \text{if } WBC_{Obs} < 3 \times 10^9/L \ (Severe\ Neutropenia) \\ 2 & \text{if } WBC_{Obs} \in [3,4) \times 10^9/L \ (Moderate\ Neutropenia) \\ 3 & \text{if } WBC_{Obs} \in [4,10] \times 10^9/L \ (Efficacy) \\ 4 & \text{if } WBC_{Obs} > 10 \times 10^9/L \ (Unefficacy) \end{cases}$$

(35)

$$HCT_{Discr}(HCT_{Obs}) = \begin{cases} 1 & \text{if } HCT_{Obs} < 45\% \ (Efficacy) \\ 2 & \text{if } HCT_{Obs} \geq 45\% \ (Unefficacy) \end{cases}$$

(36)

$PrevDose$ contains the information on the last administered givinostat dose. Possible values of $PrevDose$ are the clinically available dose levels, i.e., {50,75,100,125,150,175,200} mg/day, combined with a sign that is

negative if the treatment is interrupted due to severe toxicity and positive otherwise (Eq. 37). Instead, $PrevDose = 0$ codes the treatment interruption due to moderate toxicities ($PLT \in [75,150) \times 10^9/L$ and/or $WBC \in [3,4) \times 10^9/\mathrm{L}$ and dose triggering interruption=50mg/day, Figure 24, Panel C). This coding strategy allows to compactly store all the information about temporary interruption (i.e., triggering dose and toxicity severity) that is necessary to correctly select the resumption dose (Figure 24, panels C-D). Furthermore, $PrevDose = -1$ was a flag for the treatment start, when the initial dose has to be selected by the agent.

$$PrevDose = \begin{cases} -1 & (Initial\ State) \\ -50,-75,-100,-125,-150,-175,-200 & (Treatment\ Interruption\ due\ to\ severe\ tox.) \\ 0 & (Treatment\ Interruption\ due\ to\ moderate\ tox.) \\ 50,75,100,125,150,175,200 & (Treatment\ not\ Interrupted) \end{cases}$$

(37)

Combining all the values of $PLT_{Discr.}, WBC_{Discr.}, HCT_{Discr.}$ and each $PrevDose \neq -1$, 480 states were defined. Then, combining $PrevDose = -1$ with all $PLT_{Discr.} \geq 3, WBC \geq 3, HCT \geq 1$, 8 initial states were added. However, as all PV patients have at least one haematological parameter above the target range at the baseline [45] , the tuple $\{PLT_{Discr.} = 3, WBC_{Discr.} = 3, HCT_{Discr.} = 1, PrevDose = -1\}$ was discarded, obtaining 487 states.

4.1.2.3. QL-Agent actions

Actions were designed accounting for some safety constraints, coherently with the clinical protocol. In particular, gradual dose changes (i.e., ± 1 level with respect to the current amount) and safety treatment interruption criteria were imposed. However, the agent has a higher degree of freedom in making decisions with respect to the clinical rules (Figure 24), in order to explore and identify potentially better personalized treatments. As schematized in Table 4, QL-agent was allowed to select alternative starting doses to the standard one (i.e., $100\ \mathrm{mg/day}$). Further, in absence of severe toxicity (i.e., efficacy or inefficacy), agent was not forced to maintain the current dose level (i.e., $D =$) but could perform stepwise changes (i.e. increase/decrease by one level, $D+/D-$). Finally, a higher flexibility was given to the selection of the treatment resumption dose after severe toxicity events.

Table 4: Possible agent actions stratified by system/patient state.

$PrevDose$	QL-Agent Actions
Initial State	
-1	$50, 75,100,125,150,175,200\ [mg/day]$
States without severe toxicities: $PLT_{Discr.} \geq 3, WBC \geq 3, HCT \geq 1$	
50	$D =, D+$
75	$D =, D+, D-$

100	$D=, D+, D-$
125	$D=, D+, D-$
150	$D=, D+, D-$
175	$D=, D+, D-$
200	$D=, D-$
States with moderate/grade 1 Toxicity: $\boldsymbol{PLT_{Discr.}=2}$ and/or $\boldsymbol{WBC=2, HCT\geq 1}$	
50	0 mg/day for 14 days (temporary treatment interruption) [a]
75	$D=, D+, D-$
100	$D=, D+, D-$
125	$D=, D+, D-$
150	$D=, D+, D-$
175	$D=, D+, D-$
200	$D=, D-$
States in which treatment can be resumed following temporary interruption due to moderate/grade 2 toxicity: $\boldsymbol{PLT_{Discr.}=3}$ and $\boldsymbol{WBC=3, HCT\geq 1}$	
0	0, 50 [mg/day]
States with Severe/Grade 2 Toxicity: $\boldsymbol{PLT_{Discr.}=1}$ and/or $\boldsymbol{WBC=1, HCT\geq 1}$	
$\forall\ PrevDose$	0 mg/day for 14 days (temporary treatment interruption) [a]
States in which treatment can be resumed following temporary interruption due to severe/grade 2 toxicity: $\boldsymbol{PLT_{Discr.}=2}$ and $\boldsymbol{WBC=2, HCT\geq 1}$	
-50	0, 50 [mg/day]
-75	0, 50,75 [mg/day]
-100	0, 50,75, 100 [mg/day]
-125	0, 50,75, 100, 125 [mg/day]
-150	0, 50,75, 100, 125, 150 [mg/day]
-175	0, 50,75, 100, 125, 150, 175 [mg/day]
-200	0, 50,75, 100, 125, 150, 175, 200 [mg/day]

a: safety constraint of givinostat phase III clinical protocol

4.1.2.4. Implementation

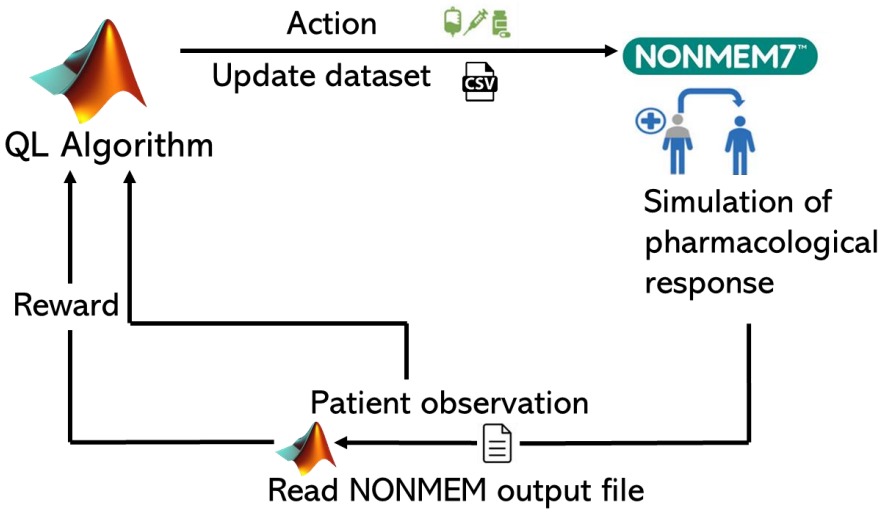


Figure 25: Schematical representation of the implemented MIRL framework to optimize givinostat precision dosing problem. In this case study, MATLAB and NONMEM were coupled to train QL algorithm and simulate patient pharmacological response, respectively.

Analogously to the erdafitinib case study presented in Chapter 2, this section introduces a novel implementation framework for MIRL [39]. Indeed, MATLAB [115] and NONMEM version 7.3.0 (ICON plc) were combined for the first time to obtain RL-based adaptive dosing strategies for givinostat. This hybrid framework was developed starting from an already available simulation platform developed to evaluate givinostat adaptive dosing protocol for the planned phase III clinical trial [45].

As illustrated in Figure 25, a simulation engine based on the available givinostat PK-PD model was embedded in the QL-framework to predict the outcome and the reward of each dosing strategy for each patient (model details in section C.2 of Appendix C). In particular, the QL algorithm is coded with MATLAB and continuously interacts with the NONMEM implementation of givinostat PK-PD model. Therefore, following QL dose selection, its PK-PD effects are inferred by running a NONMEM simulation. Then, its output file is imported in MATLAB and used to compute the reward and update patient health state.

Also in this case study, RUV was not considered in simulations, thus assuming that the status evolution of PV patient is fully described by its PK-PD model.

4.1.2.5. Learning a unique adaptive dosing protocol for the whole population with QL

The MIRL framework described in section 2.2.1 was applied to derive a set of adaptive dosing rules for givinostat suitable for all the PV patients. To this end, a unique QL-agent (QLpop-agent) was trained on a heterogeneous pool of 98 virtual PV patients (Figure 26) considering a treatment duration of 8 months, i.e., the average time to achieve a stable haematological response [45]. Then, the QLpop-agent performances were evaluated on 10 test sets, each composed by 98 virtual individuals, and benchmarked against the clinical protocol proposed for givinostat phase III study. This analysis allows to compare the robustness and the generalizability of these dosing strategies with respect to patient IIV. Both training and test sets were generated through a stratified random sampling strategy (section C.3 of Appendix C) to obtain heterogeneous virtual populations in terms of treatment response. Individual parameters and covariates of virtual patients were extracted from the parameter distributions of the givinostat population PK-PD model and the covariate distributions of subjects on which the model was originally estimated (sections C.2 and C.3 of Appendix C).

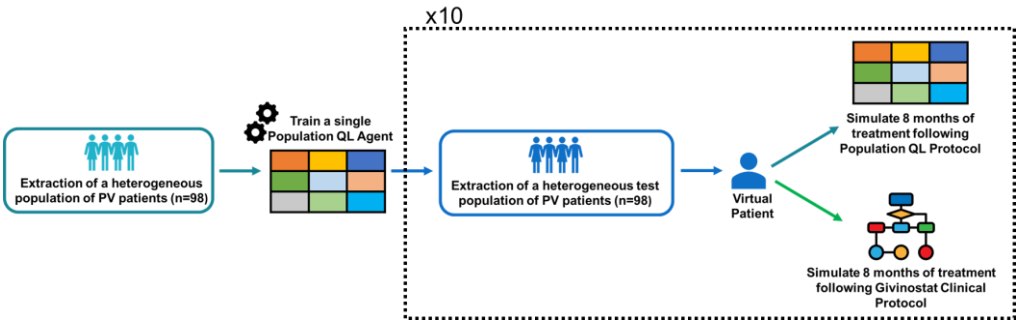


Figure 26: Schematical representation of the framework used to assess the performances of a general QL protocol (QLpop). Its robustness against inter-individual variability was compared with givinostat Phase III protocol on 10 external test sets.

4.1.2.6. Learning patient-specific adaptive dosing strategies with QL

The MIRL framework presented in section 2.2.2 was applied to tailor the adaptive dosing rules of givinostat on each specific individual by leveraging QL. Consequently, for each patient, a personal QL-agent (QLind-agent) was trained over an eight-months period and an optimal personalized dosing policy was derived. The performances of the QLind-agents were assessed on a heterogeneous population of 98 virtual patients, generated with the stratified random sampling technique detailed in section C.3 of Appendix C. Givinostat clinical protocol was used as benchmark. Furthermore, to measure the potentialities of this deeper treatment personalization, QLind-agents

were also compared with the QLpop approach. Similarly to erdafitinib case study (section 3.2.2), the comparison between QLpop and QLind agents was intentionally performed on the patient population used to train the QLpop-agent (Figure 27) to consider its most favorable conditions. As described in section 4.2.3, this framework will be extended also to use QL for deriving patient-specific adaptive dosing strategies aiming to maximize the outcome of givinostat phase III study.

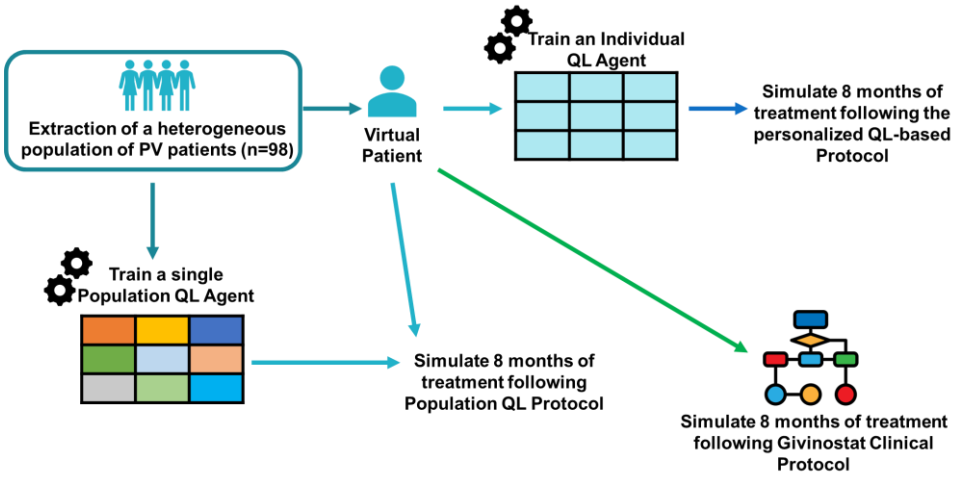


Figure 27: Schematical representation of the evaluation framework used for personal QL agents (QLind). In this case, the comparison was performed against the clinical and the QLpop agent protocols.

4.2. Results

In this section the results of the different MIRL approaches to optimized givinostat administration in PV patients will be presented. Section 4.2.1 will show the comparison between the general QL-based protocol (QLpop-agent) and the adaptive dosing rules approved for givinostat phase III trial (Figure 26). Differently, sections 4.2.2 and 4.2.3 will be focused on the results obtained with the individual-oriented QL-based workflow. In particular, the comparison between the individual QL-agents (QLind) against QLpop and the clinical protocol will be shown in section 4.2.2. Finally, in section 4.2.3 the application of individual agents to maximize the outcomes of givinostat phase III study through personalized adaptive dosing protocols will be discussed.

4.2.1. Learning a unique adaptive protocol for the whole population with QL

The QLpop-agent was trained on a virtual population of 98 PV patients (algorithm hyperparameters Table C.5 of Appendix C) with high

heterogeneity in treatment response to learn a dosing protocol suitable for the widest possible range of subjects. As detailed in section C.5 of Appendix C, the learned policy achieved low CHR rates, even in the training population. Indeed, as severe myelosuppression events were strongly penalized in the reward function (reward=0 as reported in Eq. 23), the QLpop dosing policy was biased to prefer lower doses which are unlikely to provoke severe toxicities but also enable to normalize the three hematological parameters (Figure C.6 and Figure C.**7**). Therefore, to increase the efficacy rate of the QL-based general protocol, the previous definition of reward function was revised. In particular, the too strong penalty for severe toxicities was smoothed by replacing Eq. 23 with Eq. 38, and consequently, Eqs. 25-26 with Eqs.39-40 (Figures 28 and 29).

$$Reward = Reward_{PLT} + Reward_{WBC} + Reward_{HCT}$$

(38)

$$Reward_{PLT,Obs} = \begin{cases} a \cdot e^{-0.01 \cdot |PLT_{obs} - 150|} + c & if\ PLT_{Obs} < 150 \times 10^9/L \\ 0.5 \cdot (1 - e^{-0.06 \cdot (PLT_{Obs} - 150)} + 0.5 & if\ PLT_{Obs} \in [150{,}275) \times 10^9/L \\ 0.5 \cdot \left(1 - e^{-0.06 \cdot |PLT_{Obs} - 150|}\right) + 0.5 & if\ PLT_{Obs} \in [275, 400) \times 10^9/L \\ 0.5 \cdot e^{-0.01 \cdot (PLT_{Obs} - 400)} & if\ PLT_{Obs} \geq 400\ \times 10^9/L \end{cases}$$
$$a = 0.643 \quad c = -0.143$$

(39)

$$Reward_{WBC,Obs} = \begin{cases} a \cdot e^{-0.01 \cdot |37.5 \cdot WBC_{Obs} - 150|} + c & if\ WBC_{Obs} < 4 \times 10^9/L \\ 0.5 \cdot \left(1 - e^{-0.06 \cdot (37.5 \cdot WBC_{Obs} - 150)}\right) + 0.5 & if\ WBC_{Obs} \in [4{,}7) \times 10^9/L \\ 0.5 \cdot \left(1 - e^{-0.06 \cdot |40 \cdot WBC_{Obs} - 150|}\right) + 0.5 & if\ WBC_{Obs} \in [7{,}10) \times 10^9/L \\ 0.5 \cdot e^{-0.01 \cdot (40 \cdot WBC_{Obs} - 400)} & if\ WBC_{Obs} \geq 10 \times 10^9/L \end{cases}$$
$$a = 0.643 \quad c = -0.143$$

(40)

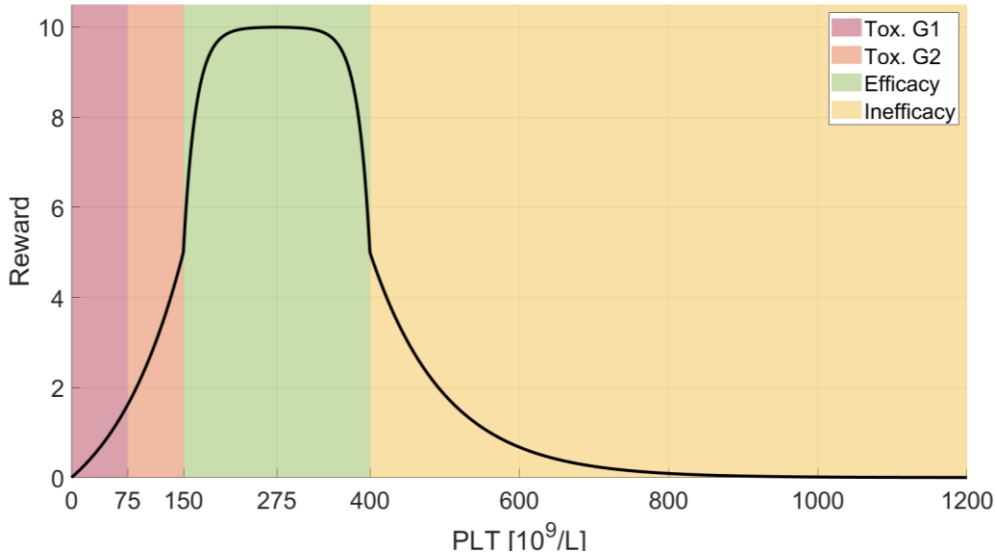


Figure 28: Plot of the smoothed version of the $\text{Reward}_{\text{PLT,Obs}}$ function introduced to improve the performances of QLpop-agent. This function evaluates the monitored values of PLT and uses an exponential decay of the reward as the observed values of PLT go below 75×10^9/L. Values are reported in the [0,10] scale accordingly to weights in Eq. 24.

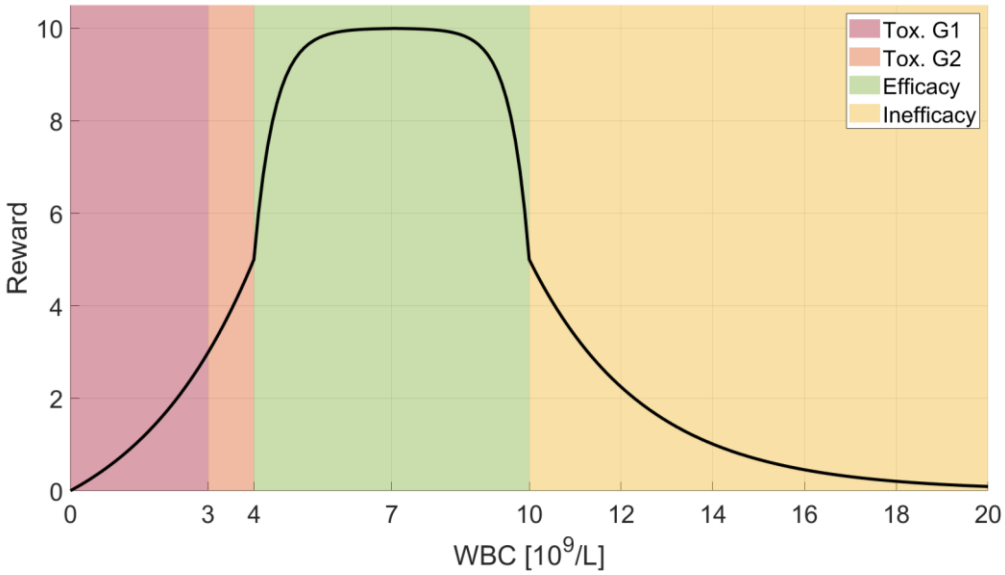


Figure 29: Plot of the smoothed version of the $Reward_{WBC,Obs}$ function introduced to improve the performances of QLpop-agent. This function evaluates the monitored values of WBC and uses an exponential decay of the reward as the observed values of WBC go below $3 \times 10^9/L$. Values are reported in the [0,10] scale accordingly to weights in Eq. 24.

As shown in section C.5 in Appendix C, with the new reward function, the CHR rate achieved by the QLpop-agent on the training population significantly increased. Then, its performances were further evaluated on 10 different test-sets and benchmarked against the phase III protocol. After 8 months of treatment, the QLpop-agent and the clinical rules induced similar rates of both single haematological parameter responses and CHR (Figure 30). The time necessary to achieve a stable CHR was comparable for the two protocols (Table 5), too. Conversely, QLpop-agent showed lower safety performances with respect to the clinical protocols as a higher percentage of severe toxicity events was observed in the 10 test sets (Table 5).

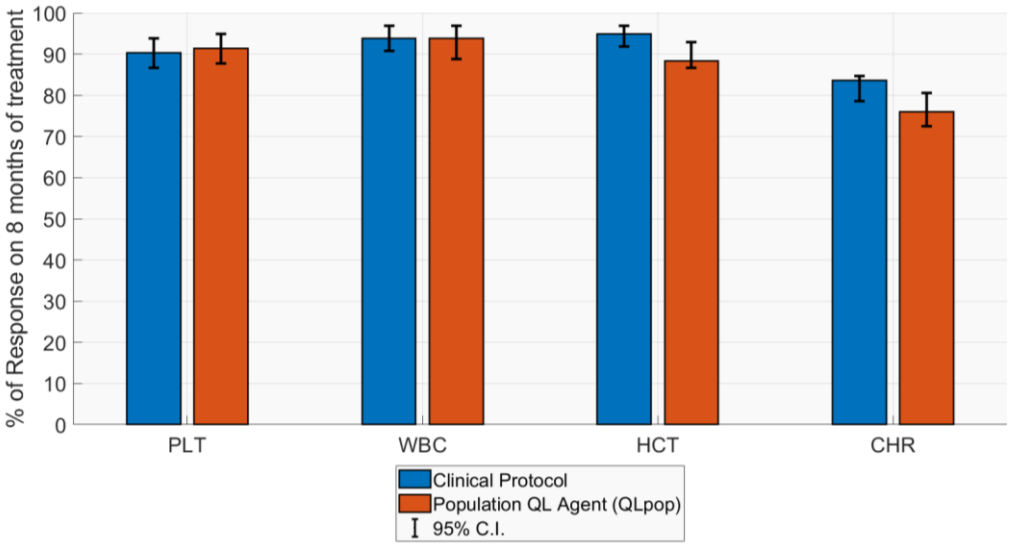


Figure 30: Comparison between givinostat Phase III protocol and QLpop dosing strategy (10 test sets) in terms of CHR and response rate for the single haematological parameter.

Table 5: Comparison between QLpop-agent and the clinical protocol proposed for givinostat Phase III study on 10 test sets composed by 98 virtual patients. Median and 95% C.I. were computed for each metric within each test set. The median of the 10 medians values, the 2.5° percentile of the 10 2.5° percentiles and the 97.5° percentile of the 97.5° percentiles are reported.

QLpop-agent	**Clinical Protocol**
Days until stable CHR [95% C.I.]	
116	110
[35, 211]	[41, 214]
% of patients with at least one severe toxicity event [95% CI][a]	
40.30	14.28
[37.75,44.89]	[11.22,17.34]
% of Days on 8-months treatment with severe Toxicity	

17.33 [3.63,46.77]	14.67 [5.5,29.11]
% of Days on 8-months treatment with $PLT \in [150, 400] \times 10^9/L$	
68.44 [24.34,100]	76.44 [32.23,100]
% of Days on 8-months treatment with WBC$\in [4, 10] \times 10^9/L$	
82.22 [38.57,100]	80.67 [37.80,100]
% of Days on 8-months treatment with HCT <45%	
92.22 [22.14,100]	90.55 [44.38,100]
% of Days on 8-months treatment with CHR	
45.33 [4.18,83.62]	52.55 [9.70,80.80]

Notes: a) 95% C.I. computed only across the 10 test sets, as for each of them it is obtained a single value

4.2.2. Learning patient-specific adaptive dosing strategies with QL

The individual-oriented QL approach was applied to the same 98-patient virtual population used to train the QLpop-agent (Figure 27). For each patient, a personal QLind-agent, characterized by the reward function described in Eqs. 23-33, was trained to specifically optimize the givinostat treatment on an 8-months period in the given subject (details on training hyperparameters in Table C.6). Then, on the same population, the performances of the QLind-agents were compared with the ones of the clinical protocol and of the previously developed QLpop-agent. As shown in Figure 31 and Table 6, the QLind-agents outperformed both the clinical and QLpop-based protocols. Indeed, after 8 months of treatment the QLind-based protocols achieved the highest rates for the single haematological parameter responses as well as for CHR (93% vs 79% and 75%). In addition, considering the whole treatment period, the QLind protocols were able to simultaneously maintain PLT, WBC and HCT in the corresponding target ranges for a higher percentage of days. QLind-agents dramatically reduced the time needed to reach a stable CHR with respect to clinical and QLpop-based rules (i.e., 47 days vs 105 and 99 days, respectively). This result is mainly due to the ability of QLind-agents in correctly selecting the highest/lowest dose (Figure 32 Figure 33) from the first cycle or in detecting patients tolerating a loading dose (Figure 34). As regards treatment safety, the individual QL-approach outperforms the two benchmarks by totally avoiding severe toxicities (Table 6 and Figure 35).

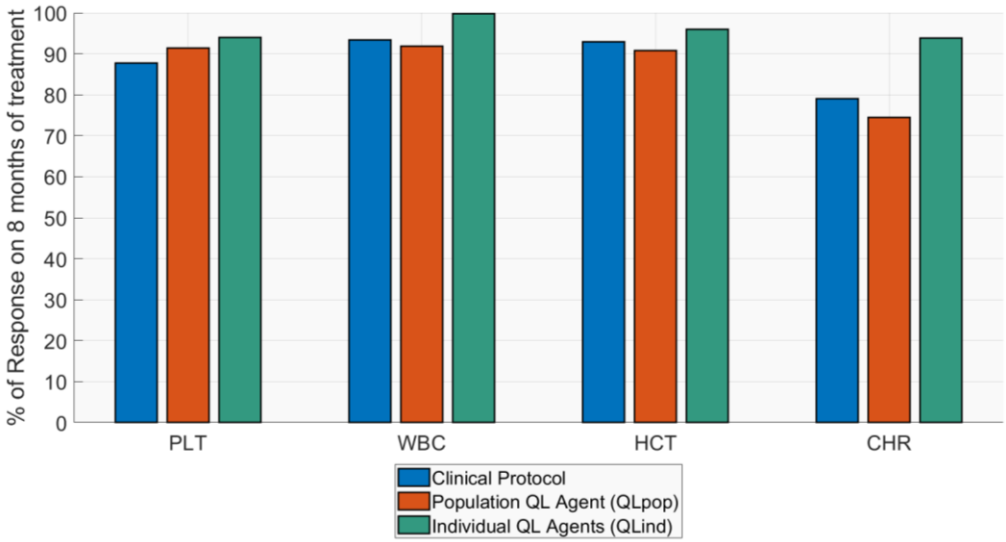


Figure 31: Comparison in terms of CHR and response rate for the single haematological parameter between QLind-agents, QLpop general dosing rules and the clinical protocol for givinostat Phase II trial. The indivudally-tailored adaptive dosing strategies used by QLind-agents outperformed the other strategies.

Table 6: Comparison between performances of QLind-agents, QLpop-agent and clinical protocol on the same virtual population of 98 PV patients. For each metric, median and 95% C.I. in the population are reported.

QLpop agent	**QLind agents**	**Clinical Protocol**
Days until stable CHR [95% C.I.]		
99 [35,211]	47 [25,147]	105 [41,214]
% of patients with at least one severe toxicity event		
43.88	0.00	12.24
% of Days on 8-months treatment with severe Toxicity		
13.78 [2.03,54.25]	0 [0,0]	17.78 [10.67,28.44]
% of Days on 8-months treatment with $PLT \in [150, 400] \times 10^9/L$		
71.78 [24.87,100]	79.78 [28.62,100]	91.56 [41.67,100]
% of Days on 8-months treatment with WBC$\in [4, 10] \times 10^9/L$		
85.33 [20.11,100]	88.44 [64.8,100]	81.33 [34.98,100]
% of Days on 8-months treatment with HCT <45%		
96.22 [20.11,100]	94.00 [46.13,100]	92.44 [44.38,100]
% of Days on 8-months treatment with CHR		
50.89 [7.51, 88.00]	75.56 [33.44,88.89]	53.33 [11.82,81.38]

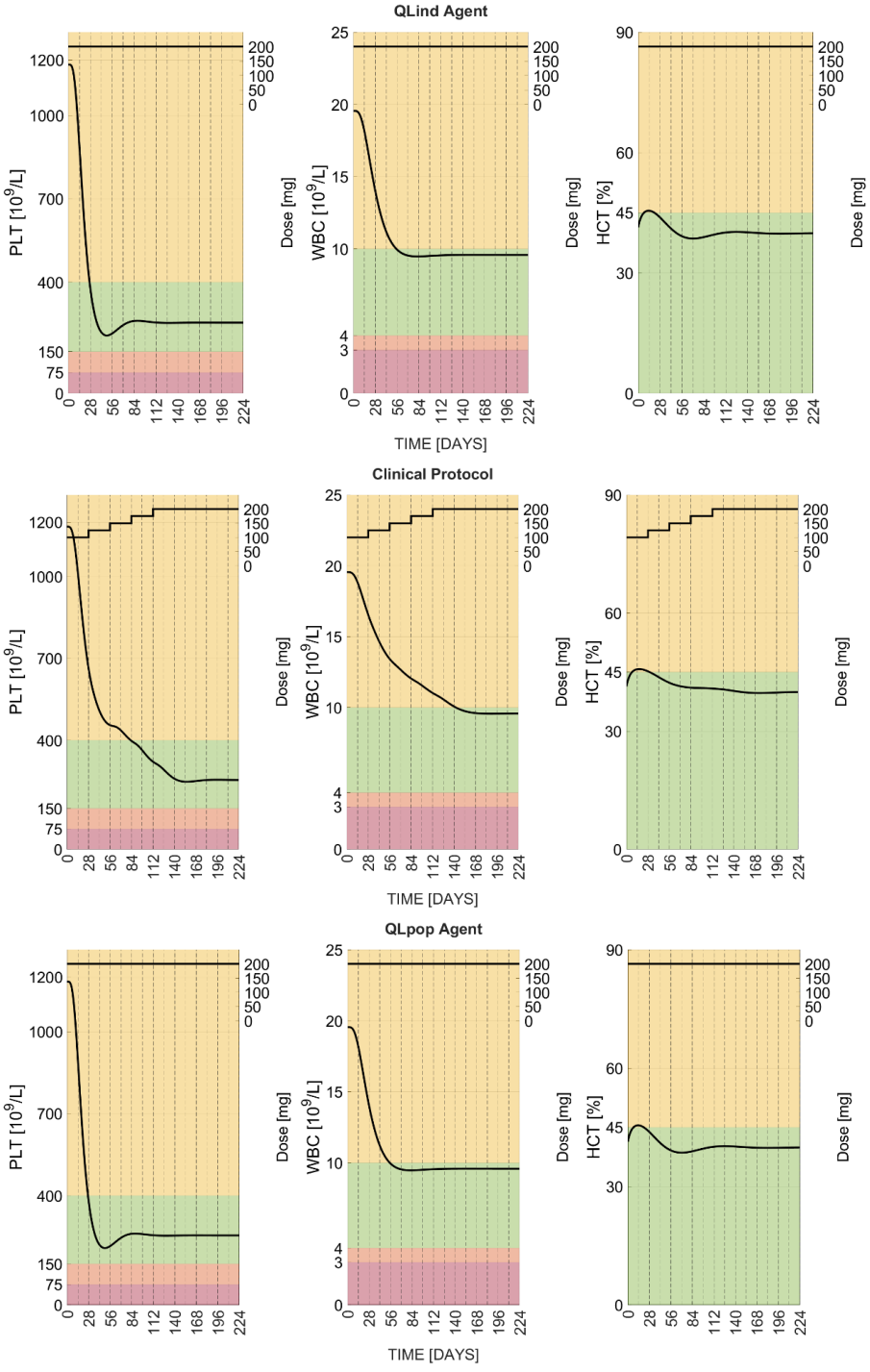


Figure 32: Example of virtual PV patient on which the individual and population QL-agent outperform the clinical protocol by choosing the maximum givinostat dose from the beginning of the treatment. This strategy allows to quickly reach the CHR. Yellow, green, orange and red shaded areas, represents inefficacy, efficacy, moderate and severe toxicity ranges of each haematological parameter.

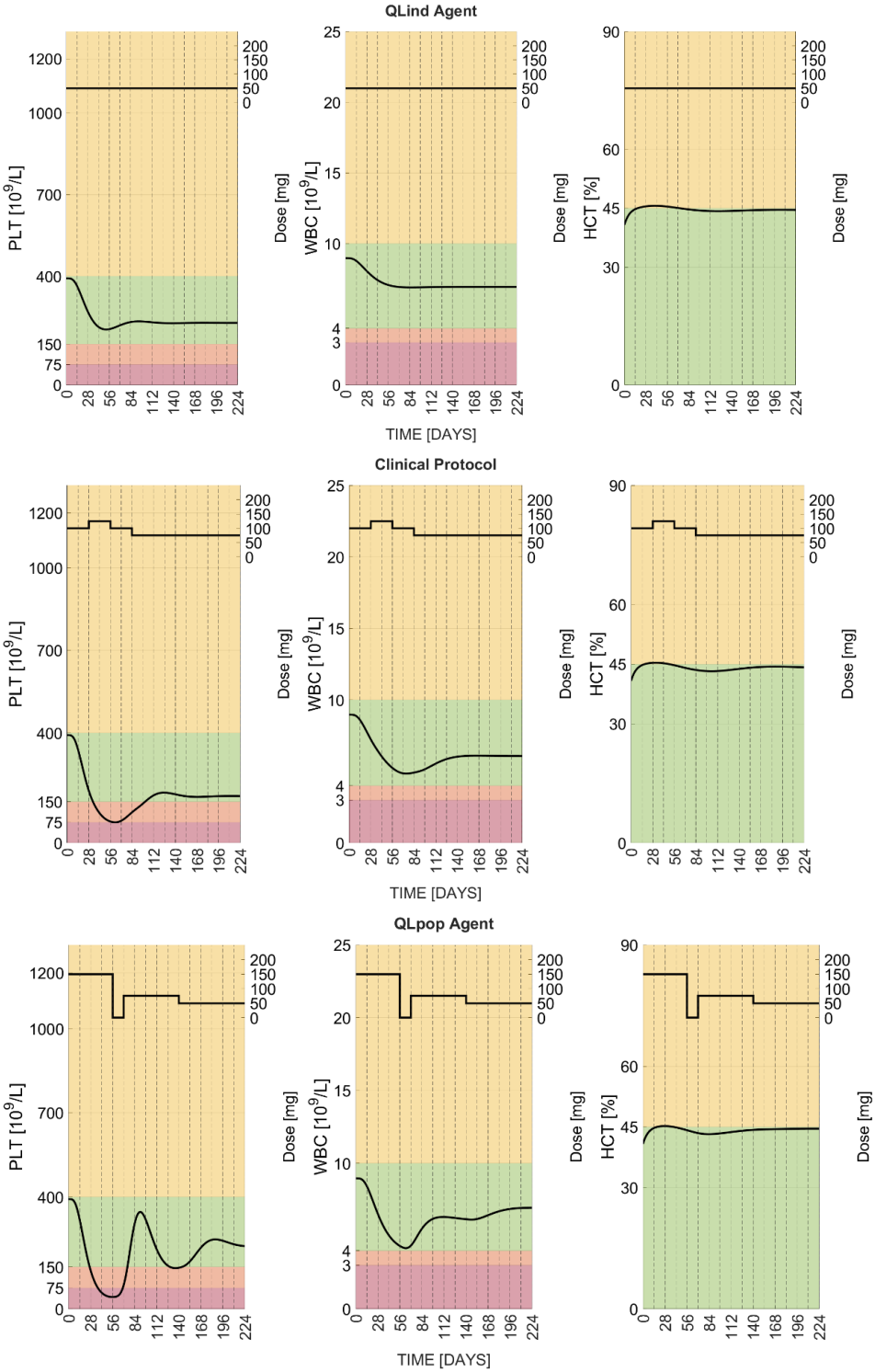


Figure 33: Example of virtual PV patient on which the QLind-agent outperforms both the QLpop-agent and clinical protocols by choosing the minimum givinostat dose from the beginning of the treatment. This strategy allows to reach the CHR avoiding toxicity events. Yellow, green, orange and red shaded areas, represents inefficacy, efficacy, moderate and severe toxicity ranges of each haematological parameter.

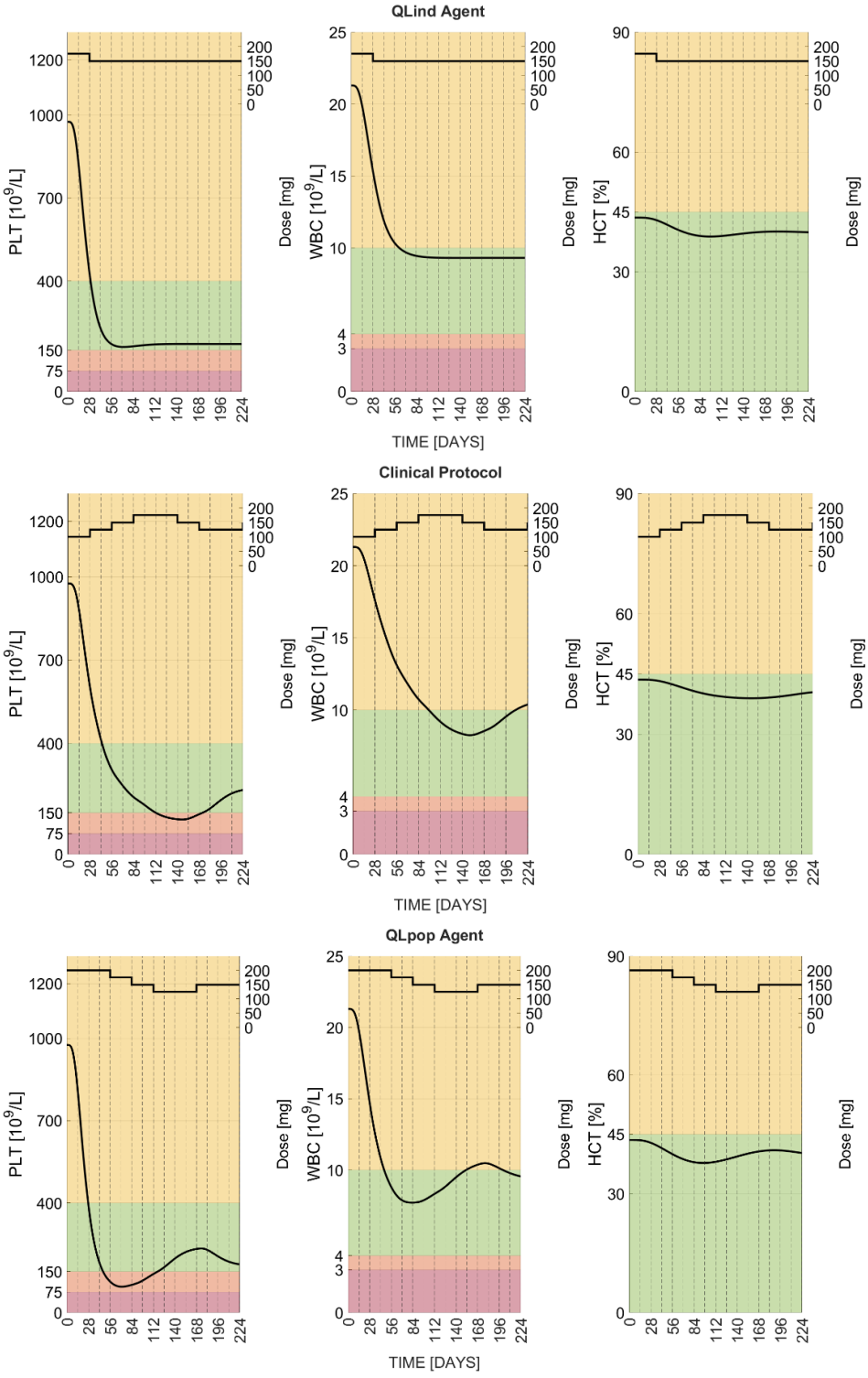


Figure 34: Example of virtual PV patient on which the QLind-agent outperforms both the QLpop-agent and clinical protocols by correctly choosing a loading dose which allows a quicker CHR. Yellow, green, orange and red shaded areas, represents inefficacy, efficacy, moderate and severe toxicity ranges of each haematological parameter

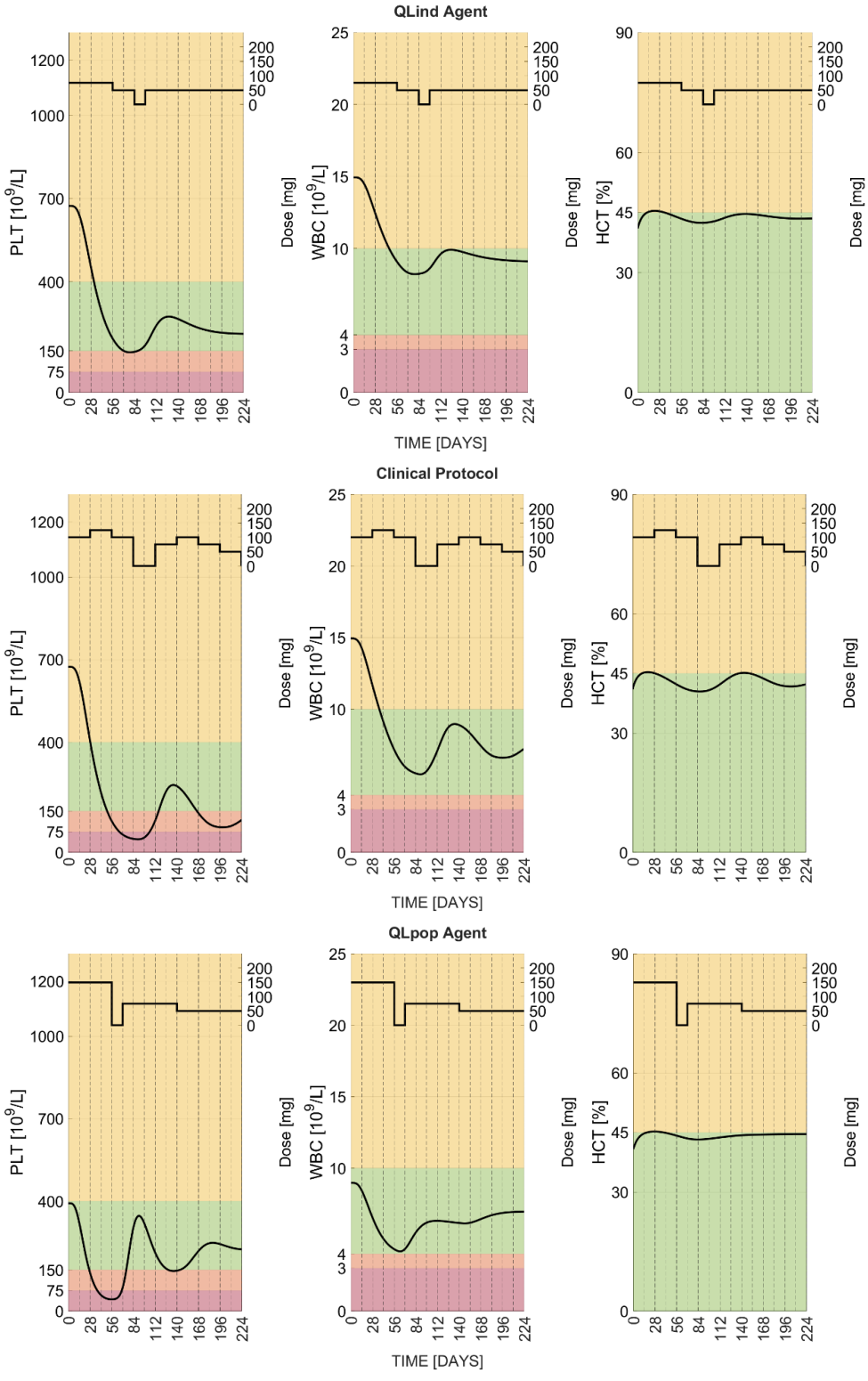


Figure 35: Example of virtual PV patient on which the QLind-agent outperforms both the QLpop-agent and clinical protocols with a dosing strategy which avoids the onset of severe toxicities. Yellow, green, orange and red shaded areas, represents inefficacy, efficacy, moderate and severe toxicity ranges of each haematological parameter.

4.2.3. Adapting patient-specific QL-based protocols to optimize givinostat phase III trial

To evaluate MIRL strategies within the drug development process, the individual-oriented QL approach was challenged to derive personalized adaptive dosing strategies maximizing the CHR rate at the eighth month of treatment. This task, in fact, represents one of the primary endpoints of the planned givinostat phase III clinical study.

The formalization of the givinostat precision dosing problem described in sections 4.1.2.1-4.1.2.3 was adopted also in this case. However, the reward function (Eqs. 23-33) adopted for QLind-agents, was properly modified to emphasize the need of achieving a CHR at the end of the treatment (i.e., 8th month). In particular, as shown in Eqs. 41 and 41, an additive bonus term was included to the reward function to renumerate more dosing actions able to induce a CHR at the end of the eighth month of treatment.

$$Reward = \begin{cases} 0 \quad if \;\; PLT_{Obs} < 75 \times 10^9/L \text{ and/or } WBC_{Obs} < 3 \times 10^9/L \\ Reward_{PLT} + Reward_{WBC} + Reward_{HCT} + \text{Bonus} \quad otherwise \end{cases}$$

(41)

$$Bonus = \begin{cases} 100 \quad if\ month = 8\ and\ CHR\ is\ achieved \\ 0 \quad otherwise. \end{cases}$$

(42)

Individual QL-agents with bonus term (QLind-bonus agents) were evaluated on the same virtual population used for the QLind -agents (i.e., those without bonus in the reward function). As shown in Table 7, QLind-bonus agents were able to achieve the desired result, that is inducing a CHR in all the patients at the end of the treatment period. However, to ensure a CHR at that specific endpoint (i.e., end of eight month), QLind-bonus agents slightly reduced treatment efficacy in the other months. Indeed, the QLind-bonus agent policies led to a shorter permanence of PLT, WBC and HCT in the efficacy ranges than the QLind-agents. Figure 36 better clarifies this difference between QLind (Panel A) and QLind-bonus (Panel B) agents. In particular, for this patient, QLind-bonus agent preferred a dosing strategy leading to a longer PLT moderate toxicity (orange shaded area) in order to obtain CHR at the eighth month. Conversely, QLind-agent, maximized the permanence in the target range (green shaded area) though CHR was not achieved at the eighth month.

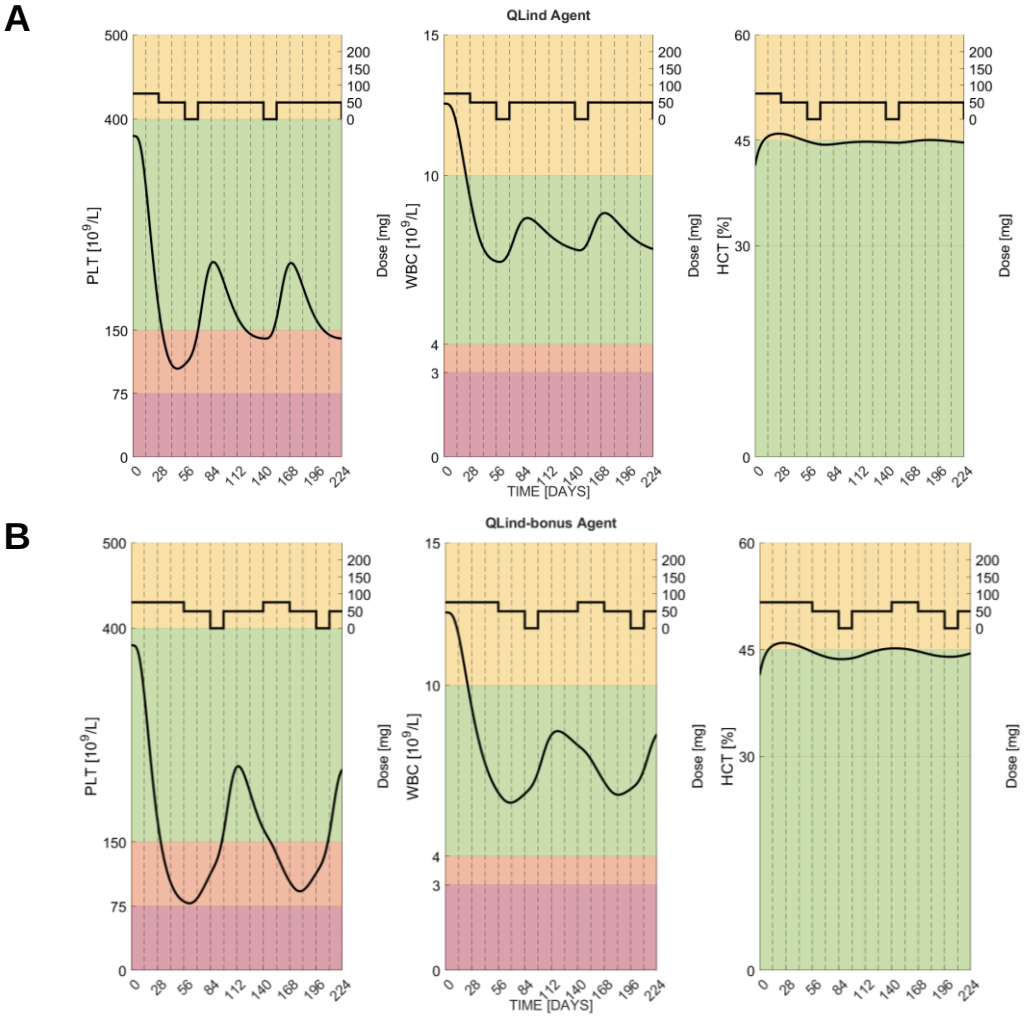


Figure 36: An example of how the introduction of bonus term in the reward function affects QL treatment personalization. For the same patient, Panel A shows the dosing strategy of QLind-agent (i.e., without bonus), conversely, Panel B illustrates the strategy proposed by QLind-bonus agent. Yellow, green, orange and red shaded areas, represents inefficacy, efficacy, moderate and severe toxicity ranges of each haematological parameter.

Table 7: Comparison between QLind-agents and QLind-bonus-agents on the same virtual population of 98 patients. For each metric, the median and 95% C.I. in the population are reported.

QLind-agents	**QLind-bonus agents**
% of Response on 8 months of treatment – PLT	
96.90	100
% of Response on 8 months of treatment – WBC	
100	100
% of Response on 8 months of treatment – HCT	
95.90	100
% of Response on 8 months of treatment – CHR	
93.80	100

% of Days on 8-months treatment with severe Toxicity	
0	0
[0,0]	[0,0]
% of Days on 8-months treatment with $PLT \in [150, 400] \times 10^9/L$ [95% C.I]	
91.56	90.44
[41.67,100]	[43.97,100]
% of Days on 8-months treatment with WBC$\in [4, 10] \times 10^9/L$ [95% C.I]	
88.44	87.33
[64.8,100]	[61.04,100]
% of Days on 8-months treatment with HCT <45% [95% C.I]	
94.00	93.11
[46.13,100]	[50.49,100]
% of Days on 8-months treatment with CHR [95% C.I]	
75.56	73.33
[33.44,88.89]	[27.09,88.89]

4.3. Discussions

In this chapter, the potentialities of MIRL approaches to optimize multi-objective treatments were explored on the clinically-derived precision dosing scenario of PV treatment with givinostat. In particular, it represents an interesting and not trivial clinical example of dose-adaptive regimen (Figure 24), for which dose adjustments depend on multiple key parameters (i.e., PLT, WBC and HCT), that are jointly considered as efficacy and safety endpoints. In addition, the high IIV affecting the haematological parameters dynamics and the givinostat response [45] further challenges the identification of a treatment regimen that can be adequate for all PV patients. Indeed, dosing rules effective for some patients could not be optimal for other individuals exhibiting not typical behaviours. These aspects make the givinostat precision dosing problem appealing for evaluating MIRL strategies, as there is a lack of knowledge on how such a framework performs in the joint optimization of multiple biomarkers.

Therefore, the sequential decision-making process relating to givinostat treatment was formalized as MDP in order to leverage RL, more specifically QL, for its optimization. System/patient states were defined including all the relevant information to inform the action choice (i.e., dose selection). Previously administered dose and a categorization of the three biomarker levels (Eqs. 34-46) were encapsulated in the state definition. QL-agent actions were defined to allow the exploration of different dosing strategies potentially better than the clinical protocol and, simultaneously, to obtain clinically acceptable rules. To this regard, only clinically available dose levels were considered, and some safety constrains (i.e., gradual dose variation, interruption due to severe toxicities) were imposed to the QL-

agents. Differently from the clinical protocol, QL-agents had more degrees of freedom in the selection of both initial and resumption doses as well as in dose-adjustments in absence of severe toxicity. Finally, the goal of givinostat treatment was formalized in the reward function described by Eqs. 23-33, integrating the literature available clinical knowledge [45,110]. Due to the multi-objective optimization problems, three equally weighted terms, one for each hematological parameter, were introduced in the reward function.

The developed QL-framework was first applied to derive a dose-adaptive protocol for givinostat suitable for all the PV patients. To this end, as detailed in section 2.2.1, a unique QL-agent (QLpop-agent) was trained on a heterogeneous population of 98 virtual patients. This training set was generated through a stratified random sampling strategy (section C.3 of Appendix C) to fully assess QL flexibility against IIV of treatment response, a central feature in precision dosing. During the training phase, the designed reward function (Eqs.23-33) was refined by introducing smoothed penalties for severe toxicities (Eqs. 38-40 and Figures 28 and 29) to increase the efficacy of QLpop-agent dosing protocol. QLpop-agent was benchmarked against the givinostat clinical protocol on 10 test-sets, each composed by a population of 98 virtual PV patients (Figure 26). The QLpop-agent and givinostat clinical protocol showed similar efficacy, even if the QLpop-agent had a bit more difficulties than the clinical protocol in avoiding severe toxicity events (Table 5 and Figure 30). This nontrivial result is a confirmation of the powerfulness of RL algorithms that are able to learn from scratch a set of reasonable rules close to those formulated on the basis of the clinical knowledge.

However, population QL-agent as well as the clinical protocol struggled to manage the presence of three endpoint to be simultaneously optimized and the high IIV, as both strategies are not tailored on each subject. endpoint to be simultaneously optimized and the high IIV, as both strategies are not tailored on each subject. These findings suggested that givinostat treatment can benefit from a deeper personalization of its dose-adaptive protocol and that RL can be exploited for this aim. Therefore, a set of personal QL-agents (QLind-agents), one for each patient, was trained and compared with the clinical and QLpop-agent protocol. The QLind-agents were able to optimize the efficacy/safety balance of givinostat treatment, to simultaneously manage the presence of multiple endpoints as well as of the high IIV, thus outperforming both population dosing strategies (Figure 31 and Table 6).

The excellent performances demonstrated by the QLind-agents confirmed the potential benefits of adopting an individual-oriented RL-approach to delivery precision dosing in a clinical setting. However, an RL-based personalization of the treatment could potentially improve also the drug development process, by avoiding attrition due to wrong dosage [72]. As described in section 4.2.3, QLind-agents were adapted to optimize the percentage of CHR achieved at the eight months of treatment, which is one of the primary endpoints of a planned phase III study [45]. To this aim, the reward function was modified to inform the RL algorithm of this specific

goal. In particular, a bonus term which renumerates actions leading the patient to CHR at the eighth month (Eqs. 41 and 42) was introduced in the reward function. These novel QLind-agents (QLind-bonus agents), were able to induce a CHR after 8 months of treatment in the entire patient population, thus maximizing the success rate of the clinical study. This simple example underlines also the flexibility of RL algorithm, which can be adapted to optimize different aspects of a pharmacological treatment, if the goal is correctly formalized through a suitable reward function.

Although the promising results in delivering precision dosing, there are some limitations. First, from an implementation perspective, the hybrid framework coupling MATLAB and NONMEM required long computational times, at least in the implementation here proposed. In particular, the adoption of NONMEM as simulation tool for patient pharmacological response (Panel B of Figure 1) coupled with MATLAB is suboptimal. The simultaneous training of 98 QLind-agents and of a QLpop-agent took 26 and 52 days on a Linux machine with 8 i7 Intel® cores (3.6 GHz of clock frequency), respectively.

Second, the limitations already discussed for the erdafitinib case study (section 3.3) are present also here. Briefly, it was hypothesized that the PK-PD model is the digital-twin of the patient and that it well describes the givinostat pharmacological response (RUV was not considered in simulations). Moreover, it was hypothesized that the digital twin of each patient PV patients was known from the beginning of the treatment (i.e., all the individual parameter of the model are known without uncertainty). Actually, individual parameter estimation requires individual data, thereby necessitating real-time monitoring during the treatment. The simplification of the real word scenario done by these assumptions was necessary to obtain comprehensible results in presence of a complex scenario with three biomarkers to control. Novel strategies to address these issues will be illustrated in Chapter 6.

In conclusion, this works highlights the powerfulness of the novel patient-centric MIRL approach also in a complicated scenario with a joint optimization of different efficacy/toxicity biomarkers. Further investigation on more complicated case studies derived from clinical practice will be performed in the next chapter.

Chapter 5

Optimization of short- and long-term outcomes of co-administered drugs. Application of the RL/PK-PD framework to axitinib/anti-hypertensive treatment in advanced renal cancer

Most of the literature applications of MIRL for precision dosing are focused on the optimization of efficacy/safety biomarkers in the setting of monotherapies [39,73,75,83,85–89,92]. However, especially in oncology, the real clinical scenario is more complicated as several drugs can be simultaneously administered to both increase the anticancer effect and compensate for the onset of treatment-related adverse events (AEs). Therefore, there is a lack of knowledge on how MIRL can perform in case of co-administrations [74]. Furthermore, although adaptive dosing strategies of anticancer drugs are generally based on the monitoring of short-term biomarkers, long-term outcomes such as overall survival (OS) or progression free survival (PFS) remain the primary endpoints [74]. Consequently, it could be of great value investigating the potentialities of MIRL paradigm to directly optimize both short- and long-term treatment outcomes.

The goal of this chapter is to explore and evaluate the performances of the personalized MIRL-based strategy on a precision dosing problem involving the co-administration of two drugs. In particular, MIRL is used to tailor an adaptive dosing strategy at patient level by optimizing first only short-term outcomes and then, short- and long-term outcomes together. Therefore, this study, for the first time, provides a comparison of these two approaches.

To this end, the challenging precision dosing problem relating to the axitinib/anti-hypertensive co-administration in advanced renal cell cancer (RCC) was used as case-study. Axitinib (AX) administration follows an adaptive dosing protocol based on the monitoring of blood pressure (BP) which represents the main efficacy/toxicity biomarker. Hypertension is the most frequent AX-induced AE and, consequently, anti-hypertensive (AH) medications are often co-administered to control an excessive BP increase. Treatment goals are pushing AX dose up to increase tumor shrinkage and, simultaneously, maintaining BP in a target range. Despite an adaptive dosing protocol is already approved, AX pharmacological response showed a considerable IIV in both exposure and efficacy/safety endpoints, analogously to other tyrosine kinase inhibitor compounds (TKI). Consequently, MIRL can be evaluated to optimize the clinical therapeutic goals of AX-AH co-administration [116,117].

This chapter is structured as follows. First, the precision dosing workflow of AX-AH co-administration will be presented in section 51.1. Then, the attention will be shifted to its mapping within RL, and in particular QL, framework Therefore, section 5.1.2 will provide a description of the implemented QL setup and its evaluation framework. Finally, sections 5.2 and 5.3 will present and discuss the obtained results. Supplementary information of this chapter is reported in Appendix D.

5.1. Methods

5.1.1. Axitinib /anti-hypertensive treatment in advanced renal cancer

Advanced RCC is one of the most common forms of kidney cancer [118,119] and it is characterized by an overexpression of vascular endothelial growth factors receptors (VEGFR), which are involved in the tumor progression, metastasis and angiogenesis [119]. Although several first-line therapies are reported in the guidelines, they can become ineffective due to the onset of a biological resistance to treatment and a second-line treatment is necessary [120,121].

AX in monotherapy has been approved in several countries for the treatment of advanced RCC after the failure of one prior systemic therapy [122]. AX is an orally administered TKI of the VEGFR and, consequently, induces vasoconstriction (Figure 37) [123]. Therefore, BP represents a biomarker for both its efficacy and toxicity [122–124].

Indeed, several works quantitatively described the relationship between AX-induced increase of BP and efficacy endpoints. For example, in [125] it was found that patients with an increase of baseline diastolic BP (dBP) greater than 10 mmHg during the treatment, had a significantly better progression-free survival (PFS). This result on PFS was confirmed and

extended on the overall survival OS through the Cox models developed in [126].

Moreover, previous phase II and III clinical studies showed that hypertension is the most common AX-related severe adverse event (AE) [125,127,128]. Hence, a normotensive condition, defined as BP$\leq$ 140/90 mmHg without the use of more than two anti-hypertensive medications, is required before starting the treatment [117,122,125,127,128].

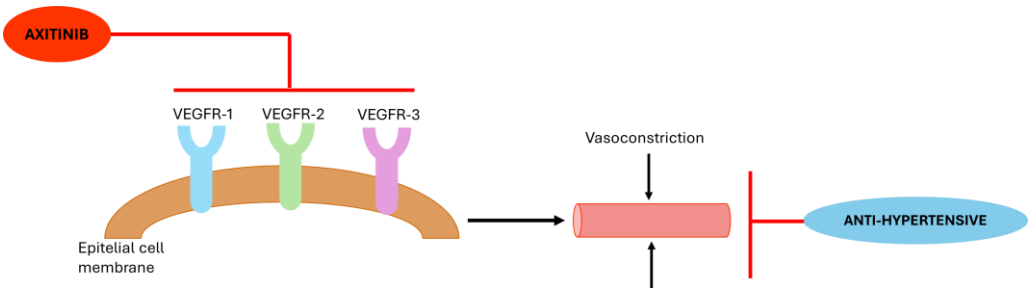


Figure 37: Schematical representation of axitinib action mechanism. AX inhibits VEGFR on the epithelial cell membrane of vascular cells. Consequently, vasoconstriction in induced and blood pression increases. Excessive BP levels are contrasted by administering anti-hypertensive medications.

AX administration follows an adaptive-dose protocol with 4-weeks cycles in which dose is adjusted based on the monitoring of BP along with other AEs [122]. In particular, the available drug levels are 2, 3, 5, 7 and 10 mg b.i.d. Treatment starts for each patient with a dose of 5 mg b.i.d. and, at the end of each cycle, the amount can be increased/decreased stepwise by one level or maintained according to criteria reported in Table 8 [122]. As concerns safety criteria, in presence of grade 4 AEs or two readings of BP>150/100 mmHg, the protocol prescribes a safety temporary interruption of the treatment until AEs improve to grade <2 and BP<150/100 mmHg. In that case, the resuming dose is one level lower than that provoking treatment interruption [122]. Only in presence of hypertensive crisis (BP > 180/120 mmHg) the treatment is permanently interrupted. Otherwise, it is continued until disease progression [122,125,127,128].

Table 8: Criteria for adjusting AX dose in the approved clinical protocol.

Changes in treatment schedule	Criteria
Dose increase	No AEs with grade >2, BP $\leq$ 140/90 mmHg without any anti-hypertensive medication
Dose decrease	AEs with grade >3 or two readings of BP> 150/100 mmHg despite maximal anti-hypertensive medication.

Although the presence of this adaptive dosing protocol, AX showed a considerable IIV in both exposure and efficacy/safety endpoints, analogously to other TKI compounds [116,117]. Therefore, it could be of interest to explore MIRL-based strategies in this context, which is furtherly challenging due to the concomitant AH medications.

Despite AHs are often prescribed to contrast the onset of AX-induced hypertension [125,129,130], no standardized protocol or consensus currently exists to guide the administration of AHs in this context [130]. Clinical studies have shown that various classes of AH drugs are suitable to manage AX-induced hypertension, and these medications can be tailored to the patient needs, either as monotherapy or in combination with other AHs [130]. Therefore, when prescribed during AX treatment, both the dosage and the number of AHs can be adjusted to control hypertension. The lack of a standardized clinical protocol led to introduce some assumptions on AH treatment within the RL formalization that will be discussed in sections 5.1.2 and in D.1 of Appendix D.

5.1.2. Set up of QL algorithm for axitinib/anti-hypertensive co-administration

First, the precision dosing of AX-AH co-administration was described as an MDP (see section 2.1.1). To this end, the essential components of MDP i.e., system states, agent actions and reward function, were defined based on the clinical scenario described on section 5.1.1. Given the lack of standard guidelines in AH administration, some assumptions were necessary. As detailed in sections D.1.3 and D.1.7 of Appendix D, the amount of administered AH medications was represented using the daily dose equivalent (DDE) formulation. This approach allows to express the quantity of administered AHs without the need to specify their exact number. In particular, the set of {0, 1, 2, 3, 4} represented the available levels of AH DDE (further details in section D.1.7 of Appendix D).

Since QL was used also in this case-study as RL algorithm, states and actions were defined in a discrete fashion. In particular, system/patient state are based on the dBP levels which are periodically monitored to guide the simultaneous dose adjustment of both AX and AHs (QL-agents actions).

Two different reward functions were developed and evaluated to optimize short- and long-term outcomes of this co-administration. In optimizing short-term outcomes, the reward function was defined to give higher remuneration to the dosing strategies that bring and maintain dBP in the efficacy range of [90,100) mmHg and, simultaneously, maximizing AX exposure and limiting AHs. Differently, for optimizing also long-term outcomes, the reward function was extended by adding a term that evaluates the effect of a dosing strategy on patient survival probability.

The empirical AX-AH PK-PD-OS model presented in section D.1 of Appendix D was embedded within QL algorithm to simulate the PK-PD-OS response of each virtual patient to individual QL-agent (QLind-agent) dosing strategies. Briefly, the available modeling framework describes AX-AH effects on dBP and directly links the antitumoral effect on the soluble form of VGEFR (sVEGFR) in plasma. Tumor growth inhibition, quantified by the lesions sum of the longest diameters (SLD), is linked to sVEGFR coherently with the AX mechanism of action (Figure 37). Finally, OS is regulated by the SLD time course.

Since AX treatment is continued until disease progression or unacceptable toxicities (i.e., hypertensive crisis) [122], a 2 years timeframe was considered for the treatment simulations in the QL setup.

From an implementation perspective, the workflow integrating R and Simulx presented in section 3.1.2.4 was adopted for this case study.

The following subsections will provide a comprehensive description of the QL setup adopted for AX-AH co-administration precision dosing problem and its evaluation setup.

5.1.2.1. Short-term reward function

The AX-AH co-administration aims to simultaneously achieve multiple short-term therapeutic goals: i) AX exposure has to be maximized in order to shrink tumor size; ii) the onset of hypertension has to be avoided and dBP needs to be maintained in the [90,100) mmHg range and iii) AH drugs should be administered only to compensate AX-induced hypertension. To account for all these aspects, the short-term reward (ST-Reward) function was defined as the sum of three different terms (Eq. 43):

$$ST\text{-}Reward = Reward_{dBP} + Reward_{AX} + Reward_{AH}. \quad (43)$$

The contributes with higher weight are $Reward_{dBP}$ and $Reward_{AX}$.The first evaluates the value of dBP observed at the end of each treatment cycle, dBP_{Obs}, as described in Eq. 44.

$$Reward_{dBP}(\mathrm{dBP_{Obs}}) = \begin{cases} 0.1 & if\ dBP_{Obs} < 60\ mmHg \\ 5 \cdot e^{-k_1 \cdot |dBP_{Obs} - 90|} & if\ dBP_{Obs} \in [60,90) mmHg \\ 1.2 \cdot dBP_{Obs} - 101 & if\ dBP_{Obs} \in [90,92.5) mmHg \\ 10 & if\ dBP_{Obs} \in [92.5,97.5] mmHg \\ -1.2 \cdot dBP_{Obs} + 127 & if\ dBP_{Obs} \in (97.5,100) mmHg \\ 5 \cdot e^{-k_2 \cdot (dBP_{Obs} - 100)} & if\ dBP_{Obs} \in (100,105] mmHg \\ 0.1 & if\ dBP_{Obs} > 105\ mmHg \\ k_1 = -\frac{1}{30} \cdot \ln 0.02 & k_2 = -0.2 \cdot \ln 0.02 \end{cases}$$

(44)

As illustrated in Figure 38, $Reward_{dBP}$ returns a higher reward (=10) to those dosing strategies bringing dBP in the middle of the target range (i.e., [92.5,97.5] mmHg interval). As dBP moves away from the center of this range in either direction (toward 90 mmHg or 100 mmHg), the reward decreases linearly until it reaches a value of 7. AX-AHs combinations that result in inefficacy or moderate toxicities are penalized with a reward <5, which decreases exponentially as dBP approaches 60 or 105 mmHg. Finally, the minimum reward (=0.1) is assigned when severe hypertension (dBP>105 mmHg) or hypotension (dBP<60 mmHg) is reached.

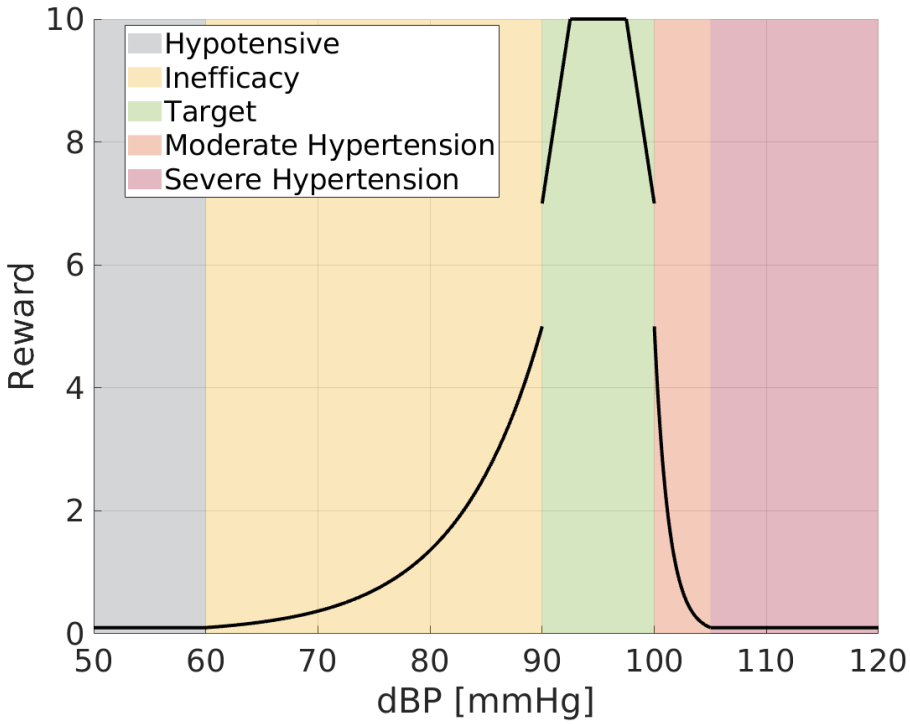


Figure 38: Graphical representation of $Reward_{dBP}$ function.

Differently, $Reward_{AX}$ (Eq. 45) is a function taking as input both the administered AX dose and the observed dBP value, dBP_{Obs}. It adds to the reward function a positive term proportional to AX dose whether $dBP_{Obs} <$ 100 mmHg. The choice of assigning a $Reward_{AX}$ >0 also in presence of $dBP_{Obs} < 90mmHg$, was made after observing a significant percentage of patients in the virtual population that cannot reach the range [90,100] mmHg even with the highest AX dose (details in section D.2 of Appendix D). Conversely, when $dBP_{Obs} \geq 100$ mmHg, $Reward_{AX} = 0$.

$$Reward_{AX}(dBP_{Obs}, Dose_{AX}) = \begin{cases} 0 & if dBP_{Obs} \geq 100\ mmHg \\ 10 & if dBP_{Obs} < 100\ mmHg \wedge Dose_{Axitinib} = 10\ mg\ b.i.d \\ 8 & if dBP_{Obs} < 100\ mmHg \wedge Dose_{Axitinib} = 7\ mg\ b.i.d \\ 6 & if dBP_{Obs} < 100\ mmHg \wedge Dose_{Axitinib} = 5\ mg\ b.i.d \\ 4 & if dBP_{Obs} < 100\ mmHg \wedge Dose_{Axitinib} = 3\ mg\ b.i.d \\ 2 & if dBP_{Obs} < 100\ mmHg \wedge Dose_{Axitinib} = 2\ mg\ b.i.d \end{cases}$$

(45)

Finally, $Bonus_{AH}$ has a lower contribute which gives a higher remuneration if lower doses of AH medications are administered when dBP is in the target range. As described in Eq. 46, this term has a significant additive contribute to the reward function only if $dBP_{Obs} \in [90,100)$ mmHg.

$$Bonus_{AH}(dBP_{Obs}, Dose_{Axitinib}) = \begin{cases} 0 & if dBP_{Obs} \notin [90,100) mmHg \\ 5 & if dBP_{Obs} \in [90,100) mmHg \wedge Dose_{AH} = 0 \\ 4 & if dBP_{Obs} \in [90,100) mmHg \wedge Dose_{AH} = 1 \\ 3 & if dBP_{Obs} \in [90,100) mmHg \wedge Dose_{AH} = 2 \\ 2 & if dBP_{Obs} \in [90,100) mmHg \wedge Dose_{AH} = 3 \\ 1 & if dBP_{Obs} \in [90,100) mmHg \wedge Dose_{AH} = 4 \end{cases}$$

(46)

5.1.2.2. Short- and long-term reward function

The reward previously presented was extended to explore the potentialities of directly integrating the optimization of patient long-term outcomes into the MIRL framework.

To this end, a new reward function, called short- and long-term reward function, (S&LT-Reward), was developed to optimize not only short-term outcomes but also patient survival probability. In particular, the proposed strategy leverages the personalization of the reward function for each patient according to individual characteristics. Therefore, each patient has an *ad-hoc* S&LT-Reward function, $S\&LT\text{-}Reward_i$, representing the balance between short-term outcomes and the individual survival probability.

As reported in Eq. 47, the individually tailored $S\&LT\text{-}Reward_i$ functions were defined from Eq. 43 by replacing the contribute of $Reward_{AX}$ with the patient-specific $Reward_{OS-AX,i}$ term.

$$S\&LT\text{-}Reward_i = Reward_{dBP} + Reward_{OS-AX,i} + Reward_{AH}$$

(47)

The $Reward_{OS-AX,i}$ functions were designed to weight the effect of each AX dose level d_j in $D = \{2, 3, 5, 7, 10\}$ mg b.i.d. on the patient-specific survival probability. Assuming that the patient parameters of AX-AHs PK-PD-OS model, $\boldsymbol{\theta_i}$, are known, for each AX dose d_j, patient survival probability at the end of a fixed-dose treatment, $S\left(t_{end}\middle|AX = d_j, \boldsymbol{\theta_i}\right)$, can be computed. As motivated in section 5.1.2, t_{end} was set equal to 2 years in this study. The $Reward_{OS-AX,i}$ was then computed for each d_j by normalizing $S\left(t_{end}\middle|AX = d_j, \boldsymbol{\theta_i}\right)$ between 0 and 10 as reported in Eq. 48.

$$Reward_{OS-AX,i}(d_j) = 10 \cdot \frac{S\left(t_{end}\middle|AX = d_j, \boldsymbol{\theta_i}\right) - S(t_{end}|AX = 0, \boldsymbol{\theta_i})}{S(t_{end}|AX = 10, \boldsymbol{\theta_i}) - S(t_{end}|AX = 0, \boldsymbol{\theta_i})}$$

(48)

In particular, the normalization uses as reference the survival probability in absence of treatment, $S(t_{end}|AX = 0, \boldsymbol{\theta_i})$, and those obtained fixing AX to the highest dose (10 mg b.i.d.), $S(t_{end}|AX = 10, \boldsymbol{\theta_i})$. Consequently, $Reward_{OS-AX,i}(d_j)$ considers the ratio of the gains of survival probability with respect to the absence of treatment between d_j and the maximum AX dose of 10 mg b.i.d. Capping Eq. 48 to 10, i.e., the maximum value of $Reward_{dBP}$ (Figure 38), is a design choice that was made to give the same importance to both maintaining dBP in its target range (short-term outcome) and optimizing OS through AX dose levels (long-term outcome).

Moreover, with Eq. 48, an excessive AX dosage should be avoided. Indeed, the highest AX doses are expected to be administered only when they can substantially bring a high gain in terms of patient survival probability.

5.1.2.3. System/patient states

Patient state was described by a tuple of four elements, $X = \{dBP_{discr}, Dose_{AX}, Dose_{AH}, FlagInterr.\}$. In particular, dBP_{discr} is the discretized dBP level observed at the end of each treatment cycle (Eq. 49). The definition of the dBP_{discr} values were based on the dBP ranges adopted by the clinical protocol described in section 5.1.1.

$$dBP_{discr} = \begin{cases} 1 \quad if dBP \in [0,90) mmHg \\ 2 \quad if dBP \in [90,100) mmHg \\ 3 \quad if dBP \in [100,105] mmHg \\ 4 \quad if dBP \in (105,120] mmHg \\ 5 \quad if dBP > 120 mmHg \end{cases}$$

(49)

$Dose_{AX}$ contains the information on the previous administered AX dose and its value can be {0, 2, 3, 5, 7, 10} mg b.i.d. Analogously, $Dose_{AH}$ represents the previous administered DDE of AH medications and, as detailed in section D.1.3, its value can be {0, 1, 2, 3, 4}. Finally, $FlagInterr.$ is a flag value which is set to 1 whether Axitinib treatment is temporary interrupted following a severe hypertension episode, 0 otherwise. This information allows to discern the states in which the anticancer therapy can be resumed after treatment interruption.

Moreover, three different initial states, S_{01}, S_{02}, S_{03}, representing patient condition before treatment (Eq. 49) were considered.

$$IntialState = \begin{cases} S_{01} & if Base_0 \in [0,70) mmHg \\ S_{02} & if Base_0 \in [70,80) mmHg \\ S_{03} & Base_0 \in [80,90) mmHg \end{cases}$$

(50)

The definition of these different and mutually exclusive initial states is based on the baseline value (i.e., pre-AX treatment) of dBP and which has a relevant impact on dBP steady-state level (see section D.1.7 of Appendix D).

5.1.2.4. QL Agent actions

Differently from the case studies presented in Chapters 3 and 4, here, QL actions are referred to a simultaneous change of both AX and AH medications. As concerns AX dosing, the set of clinical doses, {0, 2, 3, 5, 7, 10} mg b.i.d., was maintained and the actions available to QL-agents were defined accounting for specific safety constraints derived from the AX clinical protocol (section 5.1.1). In particular, gradual dose changes (i.e., ± 1 level with respect to the current amount) and safety treatment interruption criteria for severe hypertension episodes were imposed for AX. However, the agent has a higher degree of freedom in making decisions with respect to the clinical rules, in order to explore and identify potentially better personalized treatments. Indeed, following a temporary treatment interruption due to moderate hypertension, AX can be restarted at any possible dose instead of administering a 1 level lowering of the dose triggering the interruption. Furthermore, QLind-agents could start AX treatment at any dose level instead of using 5 mg b.i.d. Differently, due to the absence of a consensus/standardized administration protocol during AX treatment for AHs, QL-agents were allowed to select an AH level from {0, 1, 2, 3, 4} DDE at each treatment cycle, excluding the first one. Indeed, since all patients are supposed to be normotensive before starting the anticancer therapy, the QL-agents do not have to decide an initial dose for AH medications.

Table 9 summarizes, for each system/patient state, the actions available to QL-agents.

Table 9: Summary of the actions available to QL-agents.

System/Patient State	Possible Actions
For each state having $dBP_{discr} = 5$	Stop treatment permanently.
For each state having $dBP_{discr} = 4$ and $FlagInterr. = 0$	Stop Axitinib for at least one cycle and choose a dose for AH medications between {0,1,2,3,4}.
For each state having $dBP_{discr} \leq 2$ and $FlagInterr. = 1$	Resume Axitinib treatment with a dose in the set {2,3,5,7,10} mg b.i.d.. AH medication can be changed with a dose belonging to {0,1,2,3,4}.
For each state having $dBP_{discr} \in [3,4]$ and $FlagInterr. = 1$	Axitinib remains interrupted, only AH dose can be changed with a level belonging to {0,1,2,3,4}.
For each state having $dBP_{discr} \leq 3$ and $FlagInterr. = 0$	The previous administered dose of Axtinib can be increased/decreased stepwise by one level or maintained. AH medication can be selected between {0,1,2,3,4}.

5.1.2.5. Evaluation framework

The QL algorithm was integrated with the MIRL framework presented in section 2.2.2 to derive personalized adaptive dosing protocols for the AX-AHs co-administration. The empirical AX-AH PK-PD-OS model described in section D.1 of Appendix D was embedded in the MIRL workflow as simulation engine. A perfect model representation of the reality was assumed, consequently RUV was neglected in simulations.

In particular, individual QL-agents (QLind-agents) were trained to learn individually tailored dosing strategies with both ST and S&LT reward functions (sections 5.1.2.1-2) considering two years of treatment duration. Their performances were evaluated on the same virtual patient population in order to perform a paired comparison (i.e., the same patient treated with both dosing strategies). To this end, a heterogeneous population of 75 virtual

patients was generated by applying the stratified random sampling strategy described in section D.2. of Appendix D. Individual parameters of virtual patients were extracted from parameter distributions of the AX-AHs population PK-PD-OS model and the covariates from the distribution of subjects on which the model was estimated (sections D.1 and D.2 of Appendix D). As detailed in section D.2 of Appendix D, the adopted stratified random strategy allowed to obtain a heterogeneous virtual population in terms dBP, tumor growth and survival probability following AX-AHs treatment. Consequently, ST and S&LT reward functions were robustly evaluated with respect to IIV as a broad spectrum of treatment responses was considered. Indeed, as detailed in Table D.6 of Appendix D, the virtual population combines different levels of tumor resistance in terms of time at which AX is no longer effective with different dBP response patterns.

5.2. Results

In this section, the results of QL-based personalized adaptive dosing strategies obtained with the two rewards function previously introduced will be presented. In particular, section 5.2.1 will show the performances of QLind-agents on the virtual population when the ST-Reward function is used in the training stage. Differently, section 5.2.2 will be dedicated to a comparison between the individually tailored adaptive dosing strategies obtained by QLind-agents with both ST and S&LT reward functions.

5.2.1. Treatment personalization based on short-term reward function

The individual-oriented QL approach was first applied to the 75-patient virtual population with the ST-Reward function (Eqs. 43-45). For each subject, a personal QLind-agent was trained to individually optimize AX-AHs co-administration treatment on a 2-years period (details on training hyperparameters in Table D.7). QLind-agents were able to bring dBP in its target range for all the patients and, as shown in Figure 39, AX dose was pushed to its maximum value (10 mg b.i.d.) in all subjects starting from the 5^{th} treatment cycle. Furthermore, for some individuals well tolerating the highest AX dose level, QLind-agents were able to administer 10 mg b.i.d since the first cycle (Figure 40).

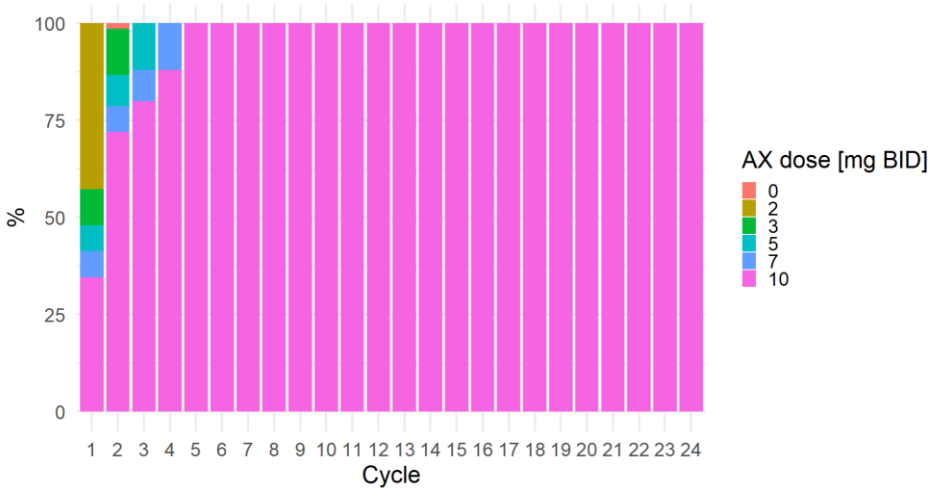


Figure 39: Distribution of AX doses administered per cycle by QLind-agents with ST-Reward function.

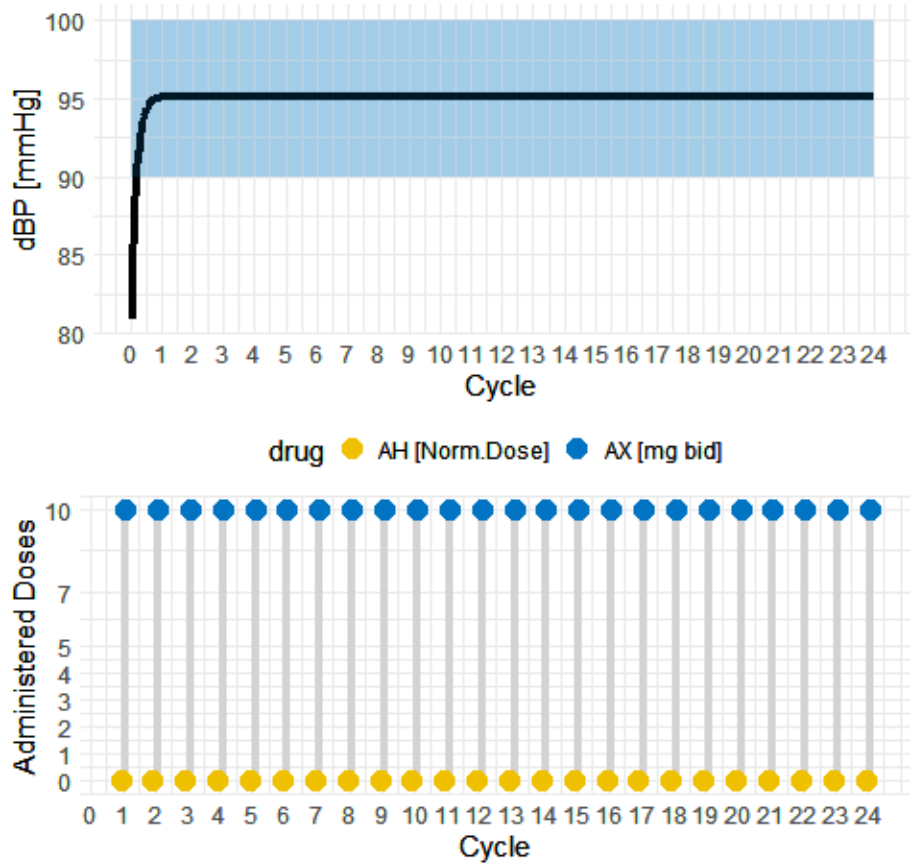


Figure 40: Example in which QLind-agent trained with ST-Reward is able to detect patient well tolerating the highest AX dose since the beginning of treatment.

Such increase of the administered AX dose did not lead to a high rate of severe hypertensions. Indeed, only one episode occurred in all the test population (1%), and it was related to a patient that would experience such severe AE for all the available AX doses, including the lowest of 2 mg b.i.d. In this scenario (Figure 42), the QLind agent started with 10 mg b.i.d. of AX, then, due to the severe toxicity, temporary interrupted the treatment (cycle 2). Following the restoration of patient normotensive condition, the treatment was continued with AX=10 mg b.i.d. and AH=3 to prevent hypertensions.

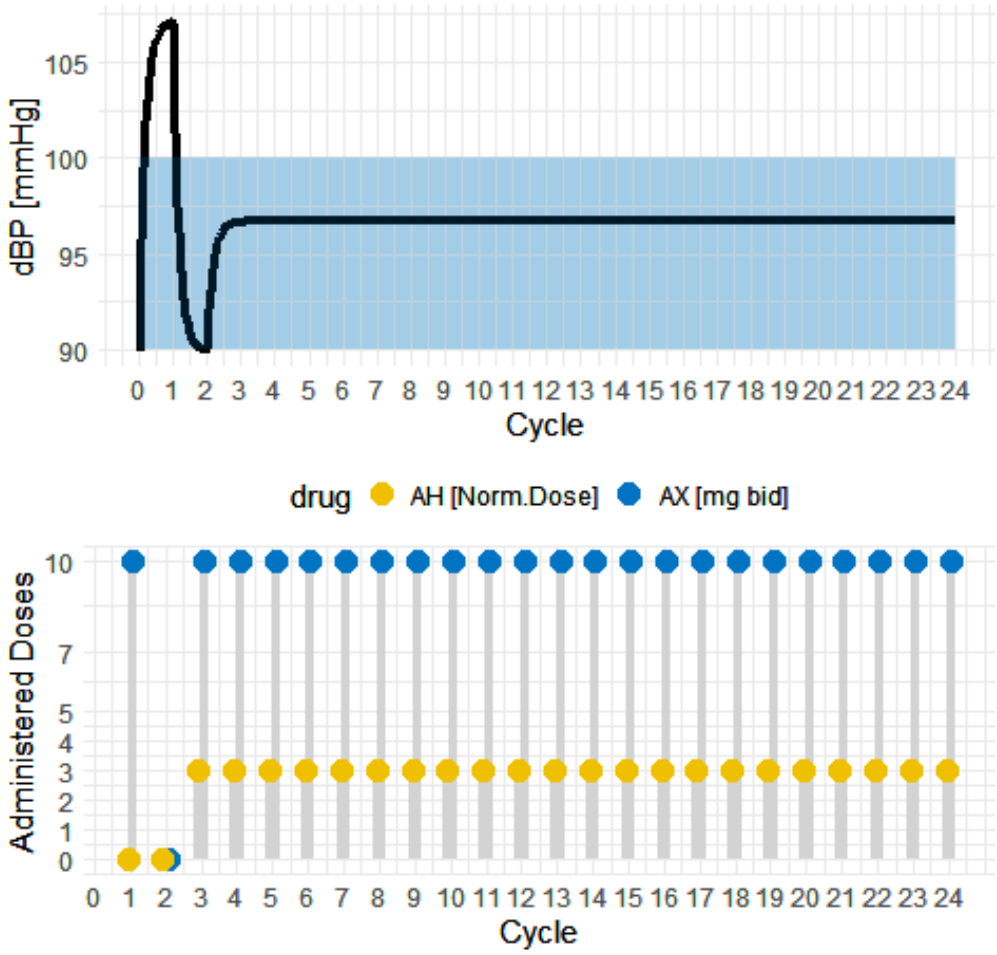


Figure 41: Example of adaptive dosing protocol learnt by QLind-agent with ST-Reward function for a patient experiencing severe toxicities for each Ax dose level.

Although AX was pushed to 10 mg b.i.d, the administered AH dosages were kept low by QLind-agents. As illustrated in Figure 42, the highest DDE of AH (=4) was never given by QLind-agents and more than the 30% of the population did not receive any AHs. When administered, AHs were mostly given with a medium dosage of 2 (30%) or 3(23%).

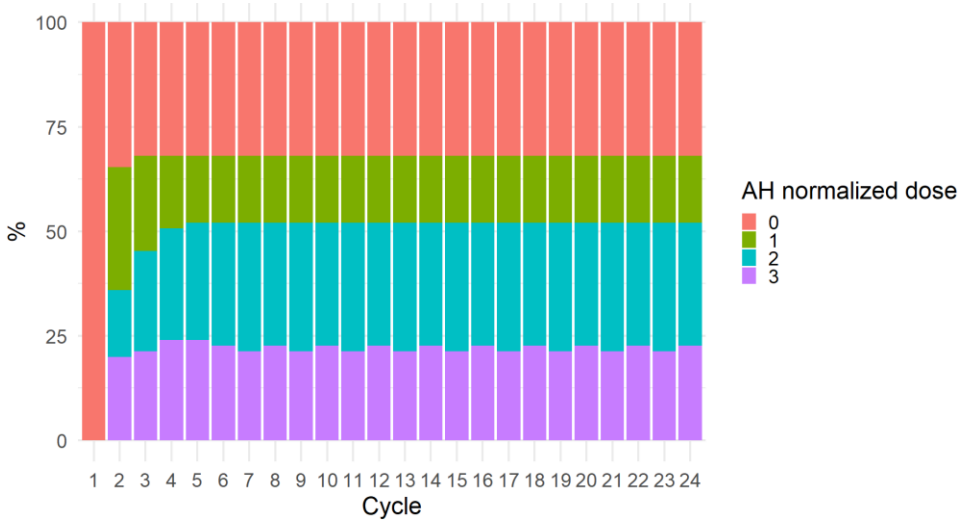


Figure 42: Distribution of AH doses administered by QLind-agents for each treatment cycle with ST-Reward function.

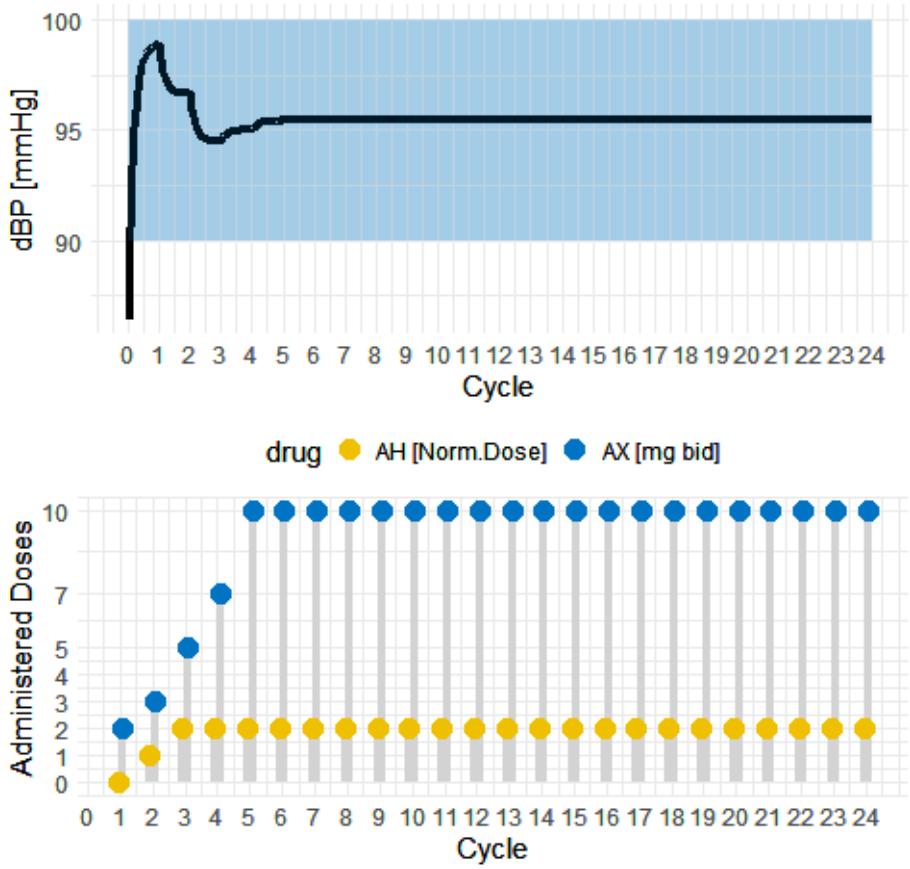


Figure 43: Example of QLind-agent trained with ST-Reward personalizing AX-AH co-administration with a joint up-titration strategy.

Interestingly, the administration frequency of 1 dose of AH was very high at the 2nd cycle (34%), then, its rate decreased. This aspect was due to dosing

strategies based on a joint up-titration of AH and AHs as for example reported in Figure 43. This patient remained in the target dBP range only with 2 mg b.i.d of AX, higher doses would provoke moderate hypertensions (details in section D.2 of Appendix D). In this scenario, QLind-agent started with the lowest AX doses and then gradually increased both the antitumoral and AHs until the minimum AH dose level compatible with 10 mg b.i.d of AX is reached.

5.2.2. Integrating long-term outcomes in the reward function

For each of the 75 patients in the virtual population, a personal QLind-agent was trained adopting the S&LT-Reward function (Eqs. 47 and 48). This approach was applied to derive individually tailored adaptive dosing strategies that optimize both short- and long-term (patient survival probability) outcomes considering 2 years of duration for AX-AHs co-treatment.

When QLind-agents trained with S&LT-Reward function were applied in the virtual population, short-term treatment outcomes were still optimized. Indeed, also in this case, severe hypertension occurred only in 1 patient (1%), the same discussed for the ST-Reward function (Figure 41). However, as illustrated in Figures 44 and 45, these results were achieved with lower AX and AH dose levels in patients. In particular, the paired AX and AH cumulative exposures over 2 years of treatment were compared in each patient. A Wilcoxon paired signed rank test ($\alpha = 0.05$, H_0= median of the differences between exposures in the two groups is equal to 0) was then performed, and statistically significant paired differences in terms of exposure at patient levels were found.

Comparing Figures 39 and 46 and Figures 42 and 47, it emerges that the QLind-agents trained with the S&LT-Reward function significantly reduced the administration frequency of both AX =10 mg b.i.d. (from 100% to 67%) and AHs=3 (from 22% to 13%). With the S&LT-Reward strategy, QLind-agents administered in some patients 7, 5, 3 mg b.i.d. of AX (10%, 9% and 8% respectively) as maintenance dose instead of pushing it to 10 mg b.i.d. Interestingly, in this scenario, also the lowest AX dose 2 of mg b.i.d. was maintained in the 5% of the patients.

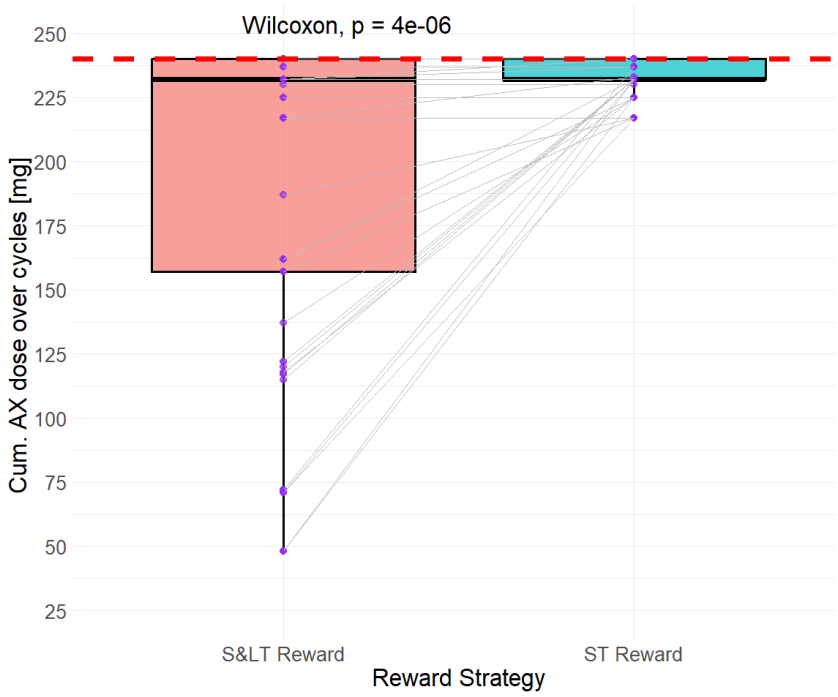


Figure 44: Paired difference of cumulative AX dose between ST and S&LT reward functions. Red dashed line represents the cumulative AX exposure with a dose of AX=10 mg b.i.d. for each cycle (=240 mg b.i.d).

Figure 45: Paired difference of cumulative AX dose between ST and S&LT reward functions. Red dashed line represents the cumulative AX exposure (=69) with a dose of AH=3 from the second to the last cycle (24^{th}).

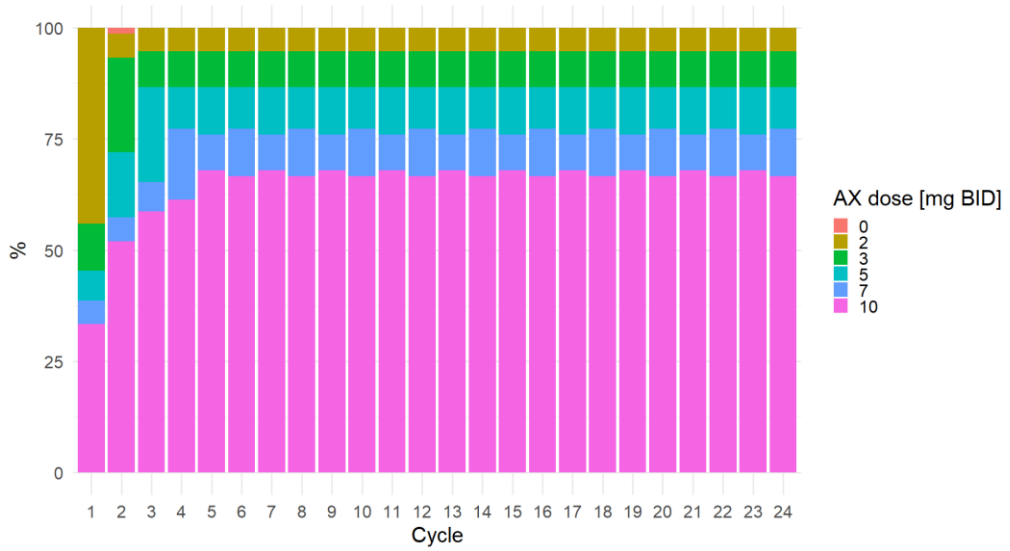


Figure 46: Distribution of AX doses per cycle administered by QLind agents with S&LT-Reward function.

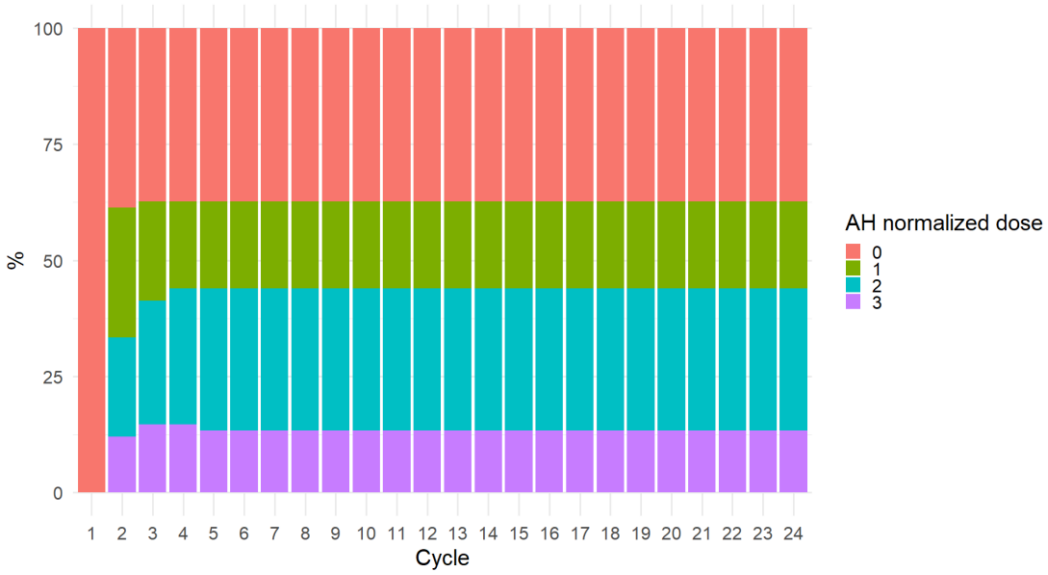


Figure 47: Distribution of AH doses per cycle administered by QLind agents with S&LT-Reward function.

Although the administered AX dose was reduced by QLind-agents trained with S&LT-Reward function, patient OS resulted similar to those obtained with the ST-Reward (Figure 48).

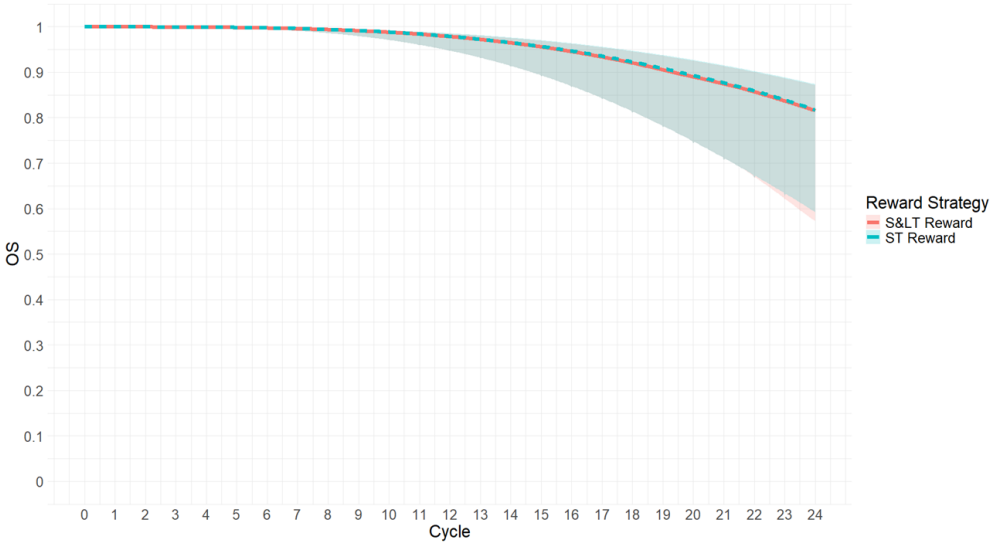


Figure 48: Comparison between OS profiles obtained with S&LT and ST reward strategies within the 75-patients virtual population

Results shown in Figures 44-48 confirm that the S&LT-Reward function allowed to define individually tailored adaptive dosing protocols that found the lowest dose of AX-AH combination able to optimize the survival at 2 years. To further demonstrate this result, some examples will be discussed below.

Figure 49 compares the profiles of dBP, tumor size (SLD) and survival probability obtained in the same patient with ST (red dashed line) and S&LT (black solid line) Reward functions.

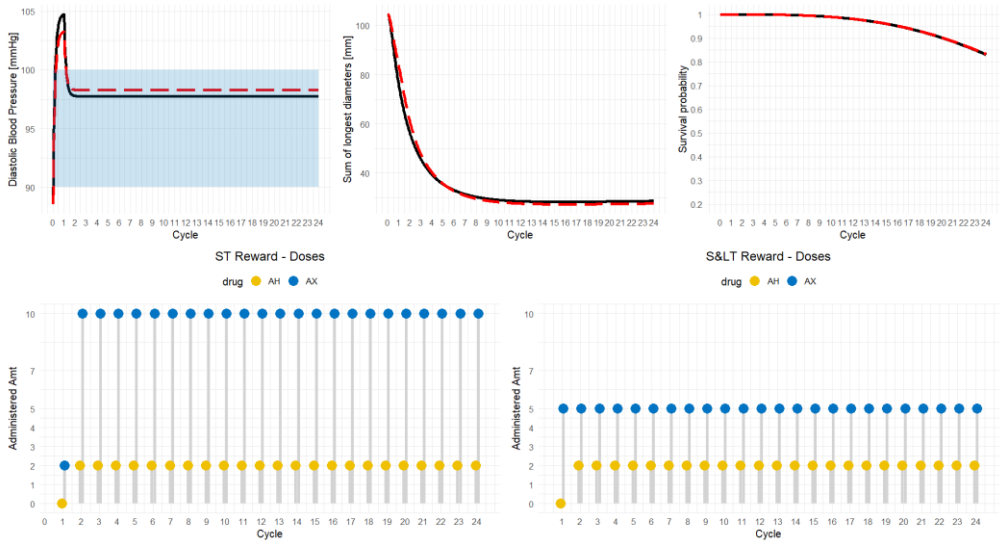


Figure 49: Comparison between the effects of the dosing strategies of QLind-agent trained with ST-Reward function (red dashed line) and S&LT-Reward function (black solid line) on dBP, tumor size (SLD) .In this patient the strategy adopted by the individual QL-agent trained with S&LT reward function limits AX to 5 mg b.i.d. instead of pushing it to 10 mg b.i.d as done when ST Reward is used.

The adaptive dosing strategy of the QLind-agent with S&LT-Reward function can be qualitatively justified by considering the weights of each AX dose assigned by $Reward_{OS-AX,i}$ (Eq. 48) according to their impact on the survival probability of this patient at 2 years from treatment beginning (Table 10). In particular, survival probability associated with AX=5 mg b.i.d is very close to those with AX=7 or 10 mg b.i.d. (gap lower than 0.01). Therefore, from a qualitative point of view, it is reasonable that administering more than 5 mg b.i.d. would not dramatically impact patient survival probability.

Table 10: Weights assigned by $Reward_{OS-AX,i}$ for each AX dose according to their impact on the survival probability at two years for the patient in Figure 49.

AX dose	$S(t_{end} = 2years \mid AX\ dose, \boldsymbol{\theta_i})$	$Reward_{OS-AX,i}$
0	0.626	0
2	0.804	8.33
3	0.817	8.97
5	0.830	9.56
7	0.836	9.83
10	0.839	10

However, by evaluating the ST Reward-based dosing strategy with S&LT-Reward, it is possible to better understand why this approach is considered suboptimal when QLind-agent is trained with the S&LT-Reward

function. The sum of discounted reward obtained by the ST-based dosing strategy was 290, three points lower than the S&LT-based one (=293). Figure 50 illustrates the differences, stratified for each of the S&LT-Reward components (i.e., AX, AH and dBP contributes, see Eq. 47), between S&LT-based and ST-based dosing strategies, when S&LT-Reward function is used to evaluate them. In particular, since both strategies give the same amount of AHs at each cycle, the difference for this component is always null. At the first cycle, the S&LT-based dosing strategy has a lower reward as it leads to a stronger moderate hypertension than the ST-based. However, starting from the second cycle, the QLind-agent trained with S&LT-Reward receives a higher remuneration for the dBP levels as they are closer to the middle of the target range [90,100) mmHg. This gain in terms of dBP is higher than the loss due to administering AX=5 mg b.i.d. instead of 10 mg b.i.d. Therefore, in presence of similar values of $Reward_{OS-AX,i}$ for AX=10 mg b.i.d and its lower doses, QLind-agents trained with S&LT-Reward are more prone to select lower doses.

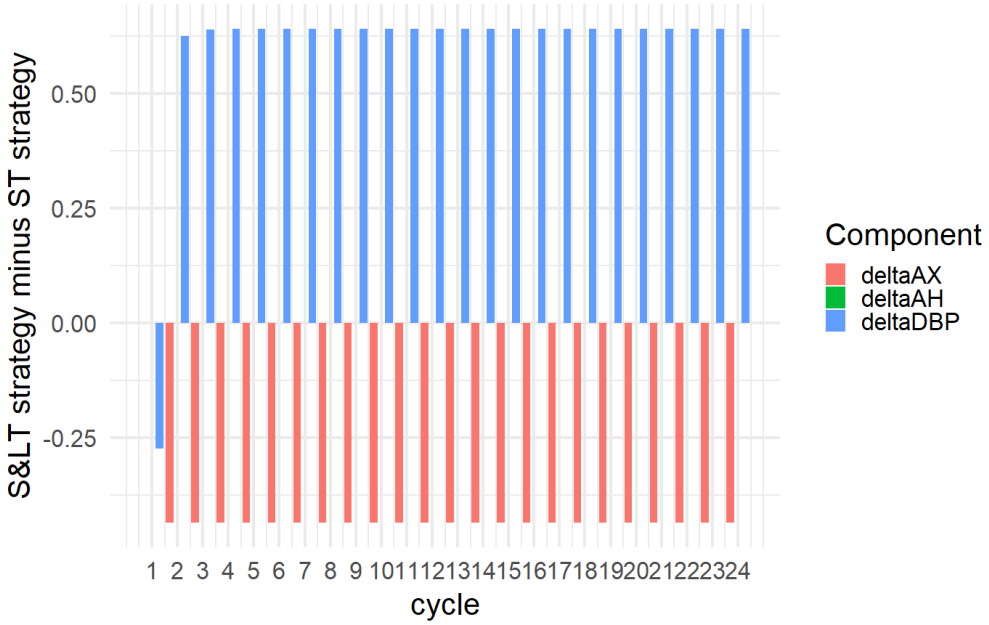


Figure 50: Stratified differences for each of the S&LT-Reward components (i.e., AX, AH, and dBP contributions, as defined in Eq. 47) between the S&LT-based and ST-based dosing strategies for the patient in Figure 49, when evaluated using the S&LT-Reward function.

Differently, when the gap between AX =10 mg b.i.d. and the lower doses is larger, the S&LT-based dosing strategy is suboptimal. To demonstrate this, the previously described analysis was repeated using the ST Reward function as matric to compare ST-based and S&LT-based adaptive dosing protocol (Figure 51). In this scenario, S&LT-based dosing strategy is no

longer the optimal one as the loss in terms of reward between administering 5 mg b.i.d. of AX instead of 10 is larger (from -0.44 to -4).

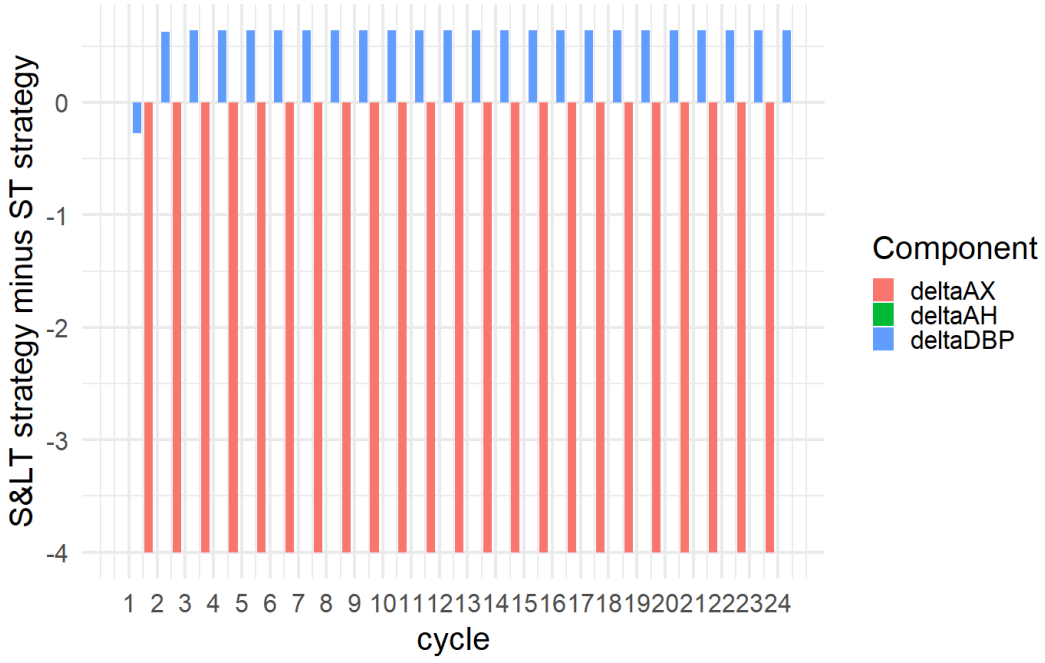


Figure 51: Stratified differences for each of the S&LT Reward components (i.e., AX, AH, and dBP contributions, as defined in Eq. 47) between the S&LT-based and ST-based dosing strategies for the patient in Figure 49, when evaluated using the ST Reward function.

A more extreme scenario in which QLind-agents with S&LT-Reward dramatically lowered AX dose is reported in Figure 52. It represents a case in which a dose of 2 mg b.i.d. is preferred instead of 10 mg b.i.d. as occurred with ST Reward function. Analogously to the previous case, the weights assigned by $Reward_{OS-AX,i}$ at each AX dose allows to qualitatively understand why this dosing strategy resulted optimal with the S&LT-Reward function. As reported in Table 11, all AX doses have almost the same impact on patient survival probability after 2 years of treatment, with a difference of 0.011 between AX=2 mg b.i.d. and AX=10 mg b.i.d.

Table 11: Weights assigned by $Reward_{OS-AX,i}$ for each AX dose according to their impact on the survival probability at two years for the patient in Figure 52.

AX dose	$S(t_{end} = 2years \mid AX\ dose, \boldsymbol{\theta_i})$	$Reward_{OS-AX,i}$
0	0.197	0
2	0.862	9.33
3	0.869	9.93
5	0.872	9.97
7	0.873	9.99
10	0.873	10

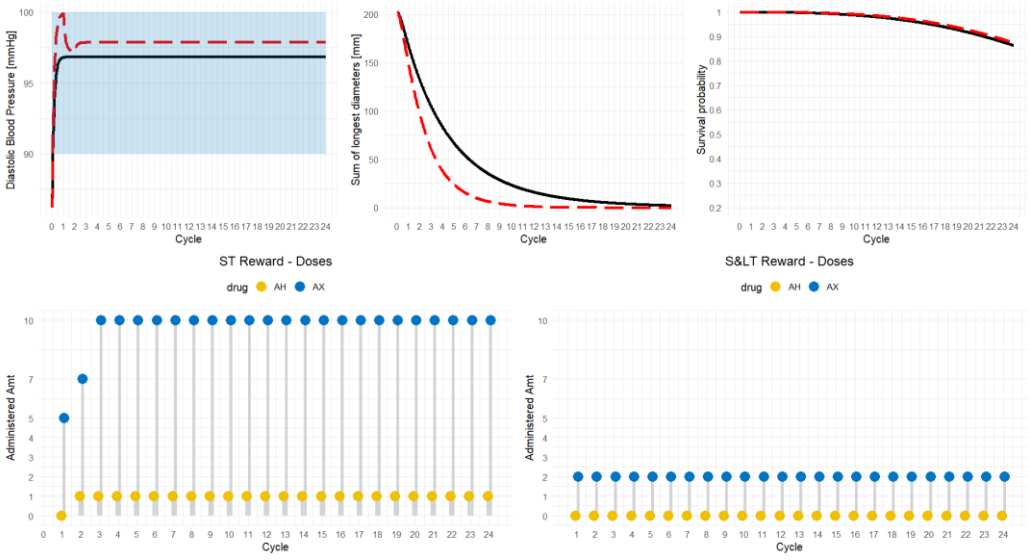


Figure 52: Comparison between the effects of the dosing strategies of QLind-agent trained with ST Reward function (red dashed line) and S&LT-Reward function (black solid line) on dBP, tumor size (SLD). In this patient the strategy adopted by the individual QL-agent trained with S&LT-Reward function limits AX to 2 mg b.i.d. instead of pushing it to 10 mg b.i.d as done when ST Reward is used.

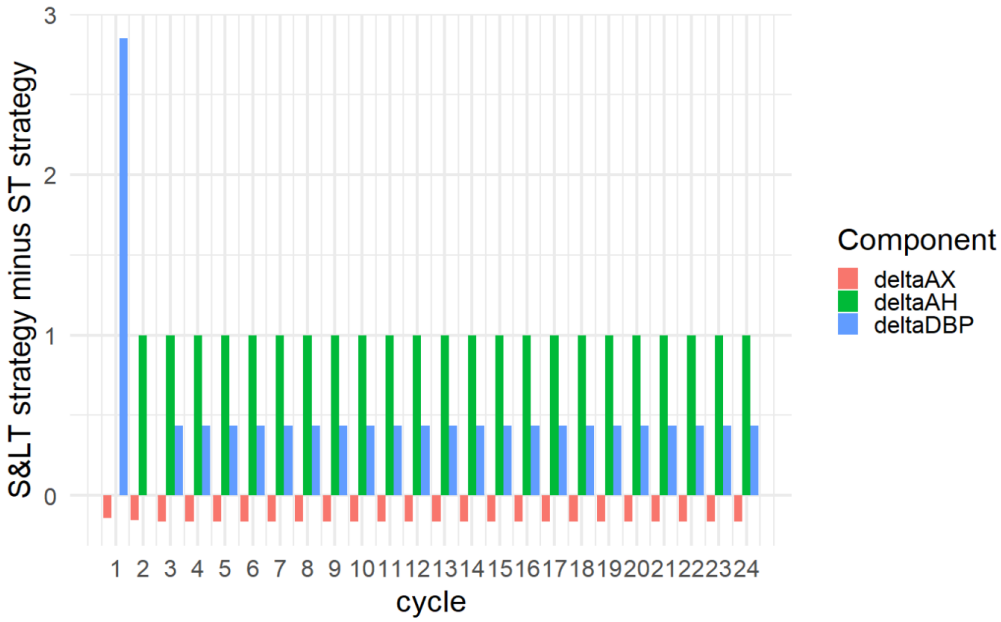


Figure 53: Stratified differences for each of the S&LT-Reward components (i.e., AX, AH, and dBP contributions, as defined in Eq. 47) between the S&LT-based and ST-based dosing strategies for the patient in Figure 52, when evaluated using the S&LT-Reward function.

When the differences for each component of the S&LT-Reward function are computed for S&LT-based and ST-based dosing strategies, the effect of having $Reward_{OS-AX,i}\,(2\ \text{mg b.i.d}) \approx Reward_{OS-AX,i}\,(10\ \text{mg b.i.d})$ is more evident (Figure 53). Indeed, by using the S&LT-Reward function, administering AX=2 mg b.i.d. is better than giving AX=10 mg b.i.d., because dBP is closer to the center of the target range and, consequently, AHs are not necessary. These aspects provide a reward gain that is much higher than the reward loss due to the use of the lowest AX dose.

Analogously to the case in Figure 49, with ST Reward the S&LT-based dosing strategy is no longer optimal due to larger reward differences between AX doses (Figure 54). Indeed, the gains in term of dBP and low AH doses are annihilated by the administration of AX=2 mg b.i.d.

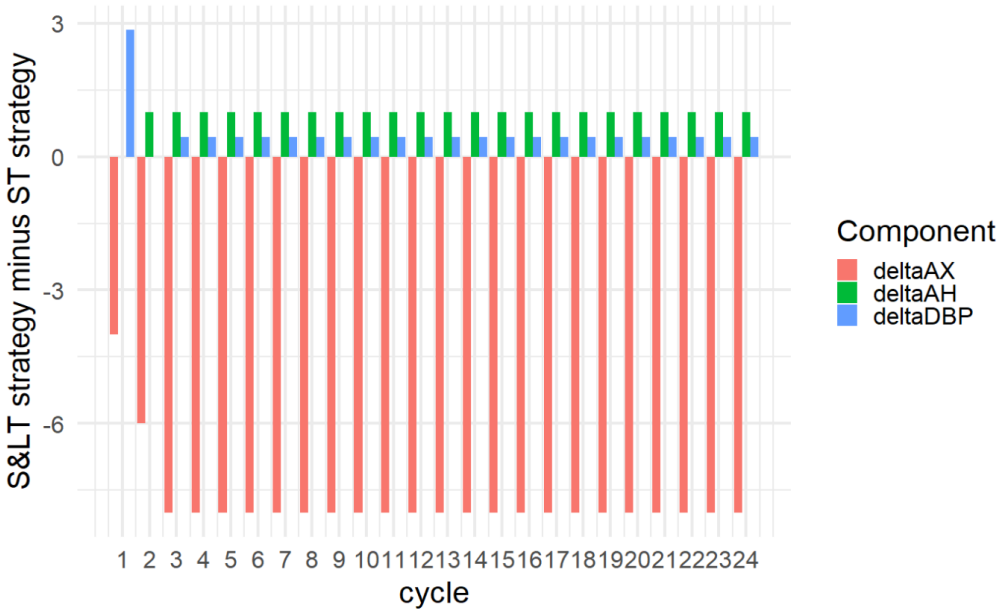


Figure 54: Stratified differences for each of the S&LT-Reward components (i.e., AX, AH, and dBP contributions, as defined in Eq. 47) between the S&LT-based and ST-based dosing strategies for the patient in Figure 52, when evaluated using the ST Reward function.

5.3. Discussions

This chapter presented the first application of the personalized MIRL framework introduced in this thesis to optimize both short- and long-term outcomes for a joint administration of drugs. To this end, AX-AH co-administration in advanced RCC patients was used as case study. In this clinical setting, AX dosage is dynamically adjusted based on the monitoring of BP which represents the primary biomarker of both AX efficacy and toxicity [116,125,127–129]. In particular, phase II and III clinical trials showed that higher AX exposure brought to higher BP increase and better survival rates. The same studies highlighted also high IIV in AX response and severe hypertensions as most common AE [117,125,127,128]. In particular, different types of AHs are often administered in monotherapy and/or combination to contrast AX-induced hypertensions [130]. Their administration does not follow a standardized protocol but they are tailored, in quantity and number, on the specific patient based on individual characteristic [130]. Consequently, despite the existence of an adaptive dose protocol for AX, it could be of interest to explore the MIRL approach to tailor simultaneously tailor AX and AH co-treatment on each single patient.

A modelling framework describing patient response, and survival probability, is central for implementing the individual oriented MIRL framework described in section 2.2.2. An already available PK-PD-OS model for AX was integrated with a PK-PD model describing the effect of

AHs on dBP during the concomitant treatment with levatinib, another anticancer compound, (see section D.1 of Appendix D) [131]. Thus, an empirical and not validated model was used in the MIRL framework to describe the AX-AH PK-PD-OS mechanisms. Although this choice represents a strong approximation, it was necessary to simulate the therapeutic problem optimization in a more realistic scenario, as AH are often administered during AX treatment [130]. Furthermore, the integration of the two modelling frameworks was supported by the fact that the introduced AH effect model was estimated on data coming from the co-administration with a compound belonging to TKI family, as AX [131].

After this preliminary step, the sequential decision-making process relating to AX-AHs precision dosing was formalized as MDP to leverage RL, more specifically QL, for its optimization. System/patient states were defined including all the relevant information necessary to inform QL-agents in jointly selecting the best AX-AH levels (i.e., agent's action). To this end, previously administered AX-AH doses and a categorization of the dBP were included in the state definition as well as a flag value representing temporary treatment interruptions.

QL-agents actions were described to represent simultaneous changes of both AX and AH medications. In particular, the available AX clinical doses, {0, 2, 3, 5, 7, 10} mg b.i.d., were used in the QL formalization. Following AX clinical protocol, safety constraints were imposed to QL-agents, such as gradual dose changes (i.e., ± 1 level with respect to the current amount) and safety treatment interruption criteria for severe hypertension episodes. On the other end, QL-agents had more degrees of freedom with respect to the clinical protocol to explore and identify potentially better personalized treatments. Indeed, QL-based protocols could customize both the initial AX dose instead of fixing it to 5 mg b.i.d. and the resumption strategies following severe hypertension. As concerns AHs, following the model calibration step described in section D.1.7, five possible daily dose equivalent (DDE) levels were defined, {0, 1, 2, 3, 4}. This approach circumvents the absence of a clinical consensus/standardized protocols as it allows to express the quantity of administered AHs without the need to specify their exact number. Moreover, at each treatment cycle QL-agents could select among all possible AH levels. Only at the beginning of treatment, the QL-based strategies were forced to start with AH=0 as AX can be started only in normotensive patients [116,117,125–129].

Finally, two different reward functions were defined and evaluated in this scenario. ST-Reward function was defined to optimize only short-term outcomes of the treatment, i.e., bringing the dBP in the [90,100) mmHg range simultaneously maximizing the AX exposure and minimizing AH administration. Similarly to what presented in Chapter 4 for givinostat treatment, this reward function was given by the sum of three components, one for each treatment goals. In particular, both dBP- and AX-related components have higher weight (0-10 reward scale) than the one related to AH (0-5 reward scale). The second reward function, S&LT-Reward, was

defined to optimize both short- and long-term (i.e., patient survival probability) outcomes. This strategy leverages the personalization of the reward function for each patient according to the individual characteristics. Therefore, an *ad-hoc* S&LT-Reward function, $S\&LT\text{ - }Reward_i$, was customized on each patient to represent the balance between short-term outcomes and the individual survival probability. In particular, each $S\&LT\text{-}Reward_i$ inherits the dBP and AH components but uses individualized weights for AX dose levels according to their impact on patient survival probability (Eq. 48). This weighting strategy was designed so that the QL-based protocol increases the AX dosage only when a higher exposure can significantly improve the patient's survival probability.

To evaluate the performances of the patient-centric MIRL approach on AX-AH co-administration precision dosing problem, a heterogeneous virtual population of 75 patients was generated. To this end, the stratified random sampling strategy described in section D.2 of Appendix D was leveraged. Therefore, for each virtual patient, both an individual QL-agent (QLind-agent) with ST-Reward and a QLind-agent with S&LT-Reward were trained to personalize AX-AH co-administration for 2 years. This strategy was adopted to perform a robust and paired comparison between ST and S&LT reward functions.

The obtained results confirmed that, also in this co-administration precision dosing scenario, the patient-tailored MIRL approach can optimize treatment outcomes with personalized adaptive dosing protocols. Individual QL-agents (QLind) successfully achieved the treatment goals on dBP with both reward functions, as all patients had this biomarker within the [90,100) mmHg target range, and severe hypertensions were dramatically reduced. Only only one episode occurred in all the test population (1%), and it was related to a patient that would experience such severe AE for all the available AX doses, including the lowest of 2 mg b.i.d. (Figure 42). The customization of the starting dose and the gradual adjustment of AX levels during each treatment played a central role in the optimization of personal dBP levels. Indeed, both reward functions were effective in identifying patients who could tolerate the highest AX dose from the beginning (Figure 40 and Figure D.3 of Appendix D), as well as those who required a gradual AX-AH up-titration to avoid severe toxicities (Figure 43 and Figure D.4 of Appendix D). Although both reward strategies obtained the same OS was obtained in the population (Figure 48), they mainly differed for the administered AX-AH dose levels, as emerged from the paired comparison of cumulative AX and AH exposures (Figures 44 and 45).

In particular, QLind-agents trained with the S&LT-Reward significantly reduced the administration frequency of the maximum AX dose (10 mg b.i.d.) compared to agents trained with the ST-Reward (67% vs 100%). Therefore, with the S&LT-Reward strategy, QLind-agents administered in some patients 7, 5, 3 mg b.i.d. of AX (10%, 9% and 8% respectively) as maintenance dose instead of pushing it to 10 mg b.i.d. Interestingly, in this scenario, also the lowest AX dose of 2 mg b.i.d. was maintained in the 5%

of the patients. As concerns AHs, both reward strategies avoided administering the highest AH level of 4 DDE. However, due to the lower AH exposure, S&LT-based strategies avoided the introduction of AH in more patients and lowered the administration frequency of AH=3 from 22% to 13%.

This result is due to the weighting strategy of AX doses in S&LT-Reward function and can be explained by considering the two examples described in Figures 49-54 and in Tables 10-11. In particular, the $Reward_{OS-AX,i}$ (Eq. 48) within S&LT-Reward function evaluates the impact of each AX dose according to patient survival probability after 2 years from treatment beginning. When the gain of survival probability between 10 mg b.i.d. and lower doses is small (e.g., 0.01), QLind-agents trained with S&LT-Reward are more likely to select a lower AX exposure. Although this results in a lower AX-related reward, this reward-loss is compensated by higher remunerations from the dBP (i.e., dBP closer to the middle of the target range) and AH (i.e., lower dose) components, which would not be achievable if the maximum AX dose of 10 mg b.i.d. were administered.

In particular, the example in Figure 52 and Table 11 highlights that a lower AX dose can achieve the same survival probability as 10 mg b.i.d. following 2 years of treatment, meaning that, in this case, increasing the AX dose does not offer a benefit for survival. However, if the goal is to quickly shrink tumor size, administering the maximum dose is more effective and this information should be encoded in the reward function.

These results remark again the centrality of the reward function in QL and, more generally, in RL. Indeed, RL algorithms blindly optimize the sum of the scores collected by the agent and returned by the reward function provided as input. Therefore, if the reward function does not accurately represent the clinical goal, an ineffective and potentially dangerous dosing strategy can be achieved. In this explorative case study, the reward function was designed based on the available clinical knowledge on AX-AH co-administration to present the possible advantages of integrating long-term outcomes (patient survival probability). However, in view of an actual clinical application, different reward functions and RL setups could be explored and designed with the aid of clinicians to ensure that all the key aspects of the AX-AH co-administration are correctly considered and formalized. Consequently, the aim of this analysis is not to establish whether S&LT-Reward is better than ST-Reward or vice versa, but rather to highlight the positive and negative aspects of these two strategies and showing how to efficiently evaluate it in the development of a MIRL framework.

Although the individual-oriented MIRL paradigm showed promising results in the optimization of both short- and long-term outcomes for multiple drug co-administrations, there are still some limitations. First, as mentioned before, the PK-PD-OS modelling framework adopted for AX-AH co-administration relies on some reasonable but not confirmed assumptions. Consequently, the obtained results cannot be properly compared to the real clinical practice. Secondly, the technical limitations already discussed for

givinostat and erdafitinib case studies (see sections 3.3 and 4.3) are still present here. Briefly, it was hypothesized that PK-PD-OS modelling framework perfectly described patient digital twin and RUV was neglected. Additionally, it was assumed that the digital twin of each patient was fully known from the start of treatment, meaning that all model parameters were considered known with no uncertainty. Under this assumption, the weights in S&LT-Reward function were computed. However, in real clinical setting, individual parameters should be estimated from patient-specific data which are collected during the monitoring performed throughout the treatment. Therefore, to practically implement S&LT-based scenario, at each monitoring step, the new patient observation should be used to update both the digital twin (i.e., patient PK-PD-OS parameters) and the weight of each AX dose in S&LT-Reward function.

However, these assumptions were made to simplify the complex scenario investigated in this Chapter and to obtain more comprehensible results. Chapter 6 will address these issues and introduce some methodological innovations to overcome them.

In conclusion, this works highlights the powerfulness of the MIRL approach to customize adaptive dosing protocols also for jointly administered drugs. It was also shown that this hybrid framework can be applied, in general, to optimize not only short-term but also short-and long-term treatment outcomes together. Further investigations will be performed to address the remaining technical issues of the framework.

Chapter 6

Overcoming Key Challenges in RL/PK-PD Framework for Clinical Integration

The applications of the MIRL approaches presented in Chapters 3-5 share some strict methodological assumptions which limit their actual implementation to support real clinical precision dosing problems.

First, it was supposed that patient digital twins used in the individually tailored MIRL paradigm (section 2.2.2) were well-characterized since the beginning of the treatment. This hypothesis translates into fully and completely knowing the individual PK-PD model parameters before the treatment starts. However, knowing individual parameters requires their estimation on patient data that generally become available during treatment monitoring.

Second, in all case studies presented so far, it was assumed that patient virtual twin is a perfect representation of the reality. Consequently, random (i.e., not mechanistically explained by the model) shifts between model predictions and real-world observations embedded in the RUV were not considered in the simulations.

The aim of this chapter is to present possible solutions to address these two issues by refining the MIRL workflow presented in section 2.2.2. To this end, two case studies derived from real precision dosing problems are leveraged.

In particular, starting from some seminal literature works [63,80,98], a Bayesian paradigm was developed to deal with the need of estimating individual PK-PD parameters. This workflow was integrated with MIRL to allow a continuous learning of model parameters from the monitored observations. As will be shown in section 3.1, this novel methodology was evaluated on a simplified version of givinostat precision dosing problem already presented in Chapter 4.

In addition, a novel extension of MIRL paradigm exploiting QL was developed to handle the stochasticity of reward function and states

transitions due to the presence of RUV representing model misspecifications. This new framework will be discussed in section 3.2 and its performances are evaluated on the precision dosing problem of vancomycin continuous infusions in intensive care unit (ICU) patients. All supplementary information to this Chapter can be found in Appendix E.

6.1. Integration of RL/PK-PD framework with Bayesian estimation

Bayesian estimation techniques are leveraged in MIPD workflow to tailor adaptive dosing protocols as schematized in Figure 2. Patient history and characteristics (e.g., covariates) and Pop-PK-PD models previously developed on large populations are used as prior information to guide the selection of the starting dose. Then, efficacy/toxicity biomarkers are monitored during treatment and observations are merged with prior information to update patient PK-PD model parameters with Bayesian estimation. Once the model is updated, it can be leveraged to perform simulations of different dosing scenarios for the next treatment cycles. Thus, the most likely dosage for the next cycle able to maintain the biomarkers in the target range is identified and used to inform dose adjustment.

This framework, starting from the seminal works in [63,80], was extended to the MIRL paradigm as illustrated in Figure 55.

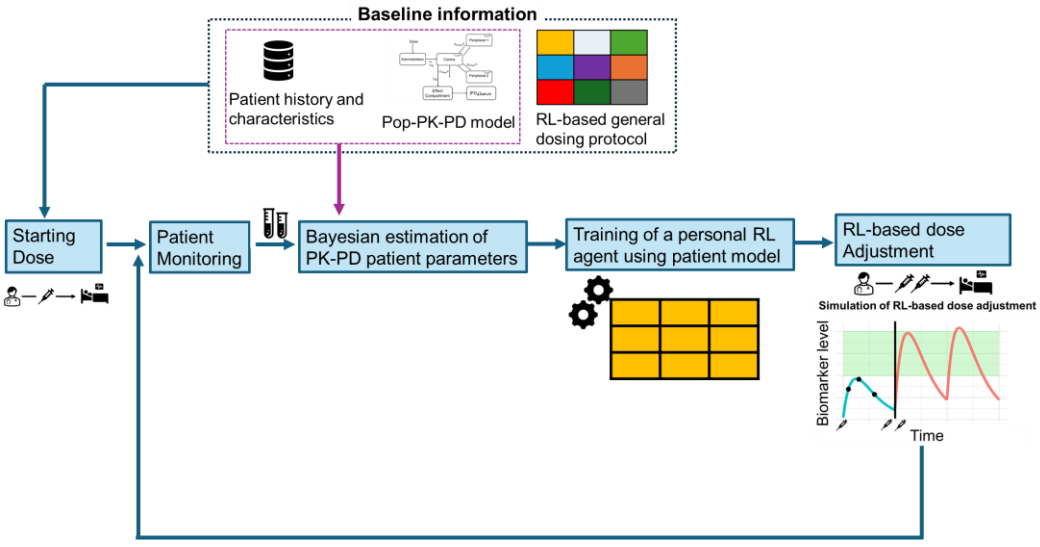


Figure 55: Extension of the MIPD precision dosing by integrating MIRL techniques.

In particular, in the MIRL context, combining *a priori* knowledge on patient population with the observed individual characteristics was translated into using a general RL-based set of rules for the selection of the initial dose (e.g., a QLpop-agent).

Then, data collected from the specific patient at each monitoring step (e.g., treatment cycle) are used to perform a Bayesian update of the individual set of PK-PD parameters. Once patient virtual replica is updated, the individual RL-agent (e.g., QLind-agent) for that specific patient can be trained to derive an individually tailored adaptive dosing protocol by using model simulations based on the latest parameter estimation. Finally, the updated RL-agent is leveraged to inform the dose selection for the next treatment cycle. This process of updating both individual PK-PD model parameters and the personalized RL-based dosing strategy repeats at each monitoring step, when further patient information is collected.

Section 6.1.1 will present an application of this Bayesian MIRL methodology on a simplified version of the givinostat case-study presented in Chapter 3.

6.1.1. Application of the Bayesian RL/PK-PD to givinostat case-stdy

The general Bayesian MIRL framework illustrated in Figure 55 was adapted as shown in Figure 56 to a simplified version of the givinostat case study presented in section 4.1.1 (details in section E.1 of Appendix E).

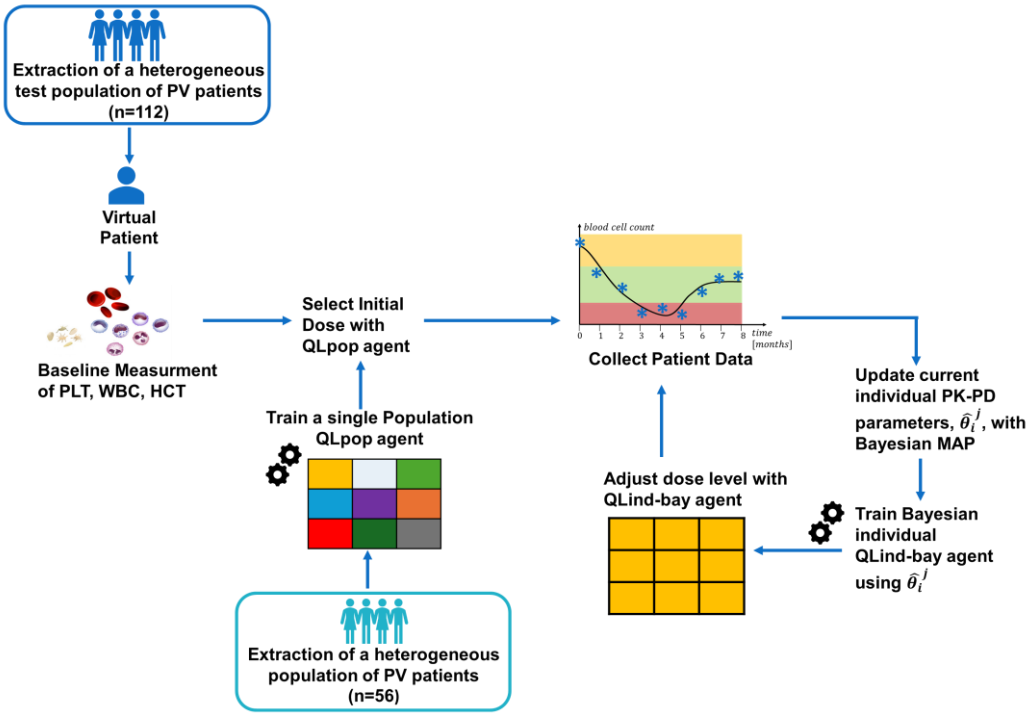


Figure 56: Schematical representation of the Bayesian MIRL workflow implemented on a simplified version of givinostat precision dosing problem.

Also in this application, QL was used as RL algorithm to personalize givinostat adaptive dosing protocol. Before treatment starts, when only baseline conditions are available, patient pre-treatment levels of PLT, WBC and HCT were fed as input to a general QL-based set of rules (QLpop-agent) to select the starting dose of the treatment. This QLpop-agent represents prior knowledge on givinostat precision dosing in PV patients and was trained on a virtual population different from the one on which the Bayesian MIRL framework was evaluated (training details on section E.3 of Appendix E).

Then, the monitored values of PLT, WBC and HCT at the 28th day (i.e., end of the first treatment cycle) were observed and used to estimate the PK-PD parameters of each patient digital twin, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\mathbf{1}}$, where superscript indicates the treatment cycle. To this end, Bayesian maximum a posteriori (MAP) [63] estimation was performed to integrate the prior information of givinostat Pop-PK-PD model parameters with the current observations of PLT, WBC and HCT. Finally, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\mathbf{1}}$ is given as input to the givinostat PK-PD model to train an individual QL-agent (QLind-bay agent) which is leveraged to select the dose for the second treatment cycle.

Therefore, at each treatment cycle j, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j-1}}$ is updated through Bayesian MAP estimation by integrating the latest PLT, WBC and HCT observations, all the previously collected ones and the givinostat Pop-PK-PD model as prior information. This updated PK-PD parameter set of patient digital twin,

$\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j}}$, is then used to refine the patient-specific adaptive dosing protocol by retraining the QLind-bay agent. In this process, the patient optimal dosing policy obtained at the previous treatment cycle is used as initial point for the new estimation. For the first training of QLind-bay agent, QLpop-agent is used as starting point. Using the QLpop-agent as a decisional set of rules for the first cycle can be considered equivalent of fixing the QLind-bay agent dosing policy to that of QLpop-agent.

This framework was evaluated to personalize the adaptive dosing protocol of givinostat on a heterogeneous cohort of 112 patients generated with the stratified random sampling strategy described in section E.2 of appendix E. Patient treatment response as well as the observed PLT, WBC and HCT values were simulated by using givinostat PK-PD model and their real PK-PD parameters, $\boldsymbol{\theta}_{\boldsymbol{i}}$.

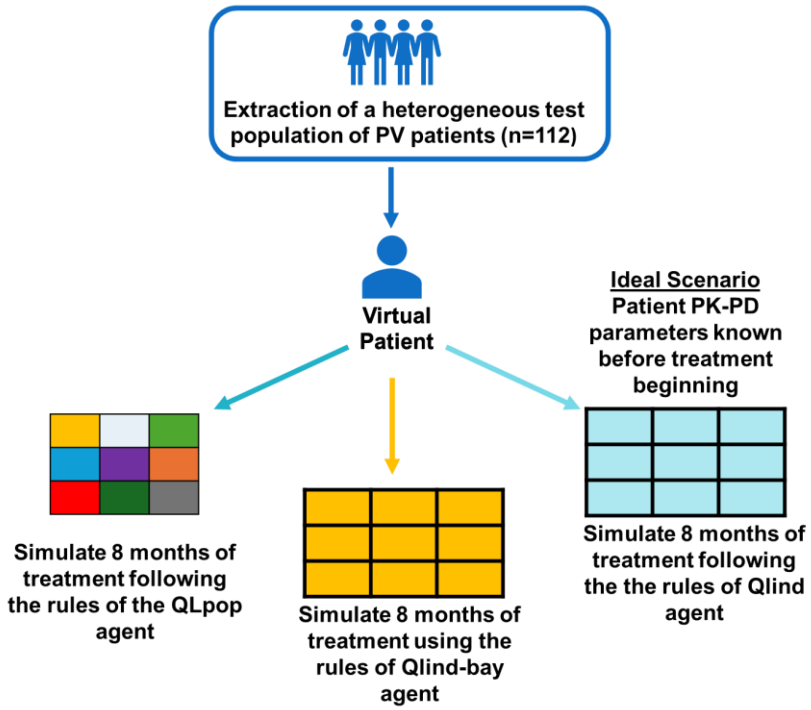


Figure 57: Evaluation framework for the proposed Bayesian MIRL approach.

As illustrated in Figure 57, the performances of the QLind-bay agent characterizing the Bayesian MIRL approach were benchmarked against both the QLpop-agent and the QLind-agents on the same virtual population. The QLpop-agent (section 2.2.1) represents the case in which a general RL-based protocol optimized for an entire patient population is used to guide patient dosing without performing individually tailored dosing strategies. Differently, the QLind-agents reflect the less realistic scenario applied in the Chapters 3, in which $\boldsymbol{\theta}_{\boldsymbol{i}}$ are fully known before the treatment begins.

From an implementation perspective, the framework integrating MATLAB and NONMEM that was applied in Chapter 3, was extended and applied also here. More specifically, Bayesian MAP estimation as well as the simulation of givinostat PK-PD model were performed in NONMEM, differently the QL algorithm was coded in MATLAB. To better understand the performances of the Bayesian MIRL approach and to avoid the simultaneous introduction of two layers of complexity in the MIRL framework, RUV was not considered in the simulations.

6.1.2. Results

Before applying the Bayesian MIRL approach, a QLpop-agent was trained to identify a set of rules for selecting givinostat initial dose according to patient baseline characteristics. QLpop-based adaptive dosing protocols reached performances similar to the ones of givinostat clinical protocol (details in section E.3 of Appendix E).

All individual QLind-bay agents were initialized to QLpop-dosing rules before treatment begins in order to use the general QL-based rules to select the starting dose. Table 12 contains the dosing rules derived by QLpop-agent and used by QLind-bay agents in the Bayesian MIRL to select the initial dose for each patient according on individual baseline characteristics.

Table 12: Optimal givinostat starting doses selected by QLpop-agent according to patient baseline levels of PLT, WBC and HCT. This set of rules was applied within Bayesian MIRL framework to personalize givinostat initial dose levels according to pre-treatment individual characteristics.

Baseline characteristic within normality range (Yes/No)			**QLpop-agent initial dose level**
PLT	WBC	HCT	
No	No	No	150 mg/day
Yes	No	No	150 mg/day
Yes	Yes	No	150 mg/day
No	Yes	No	150 mg/day
No	Yes	Yes	50 mg/day
No	No	Yes	150 mg/day
Yes	No	Yes	200 mg/day

Following the selection of the starting doses, the workflow in Figure 56 was applied for each of the 112 virtual patients in the virtual population. At each treatment cycle j, an individual set of parameters $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j}}$ was estimated through Bayesian MAP by leveraging both actual and past patient observations. Then, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j}}$ was leveraged to update the individualized adaptive dosing strategies by retraining the QLind-bay agent. As illustrated in Figure 57, QLpop-agents and individual QLind-agents were used as benchmarks to

evaluate QLind-bay agents. QLind-agents represent in this comparison the ideal scenario in which all patients PK-PD parameters are known before treatment begins.

Figure 58 and Figure 59 illustrate the response rates achieved by these three strategies at the end of each treatment cycle until the last one (8th cycle) and the distribution of virtual patient PLT, WBC and HCT profiles generated by each strategy, respectively.

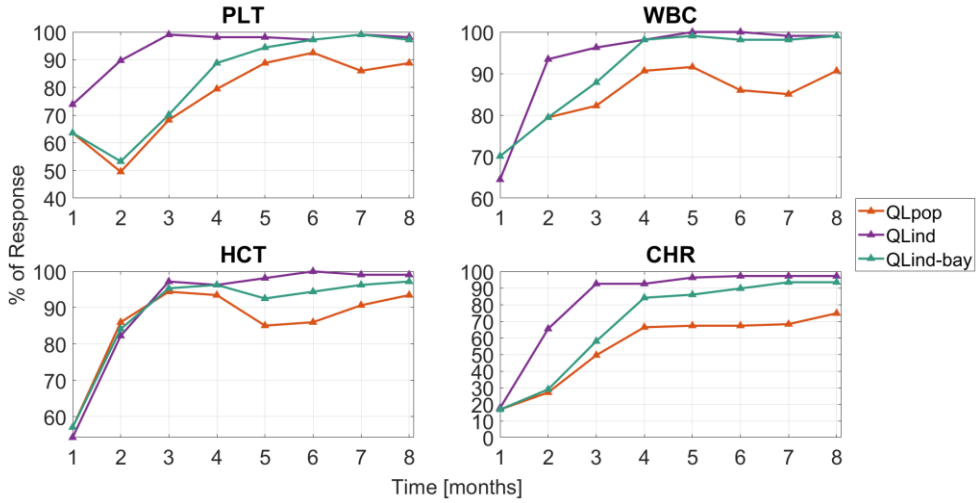
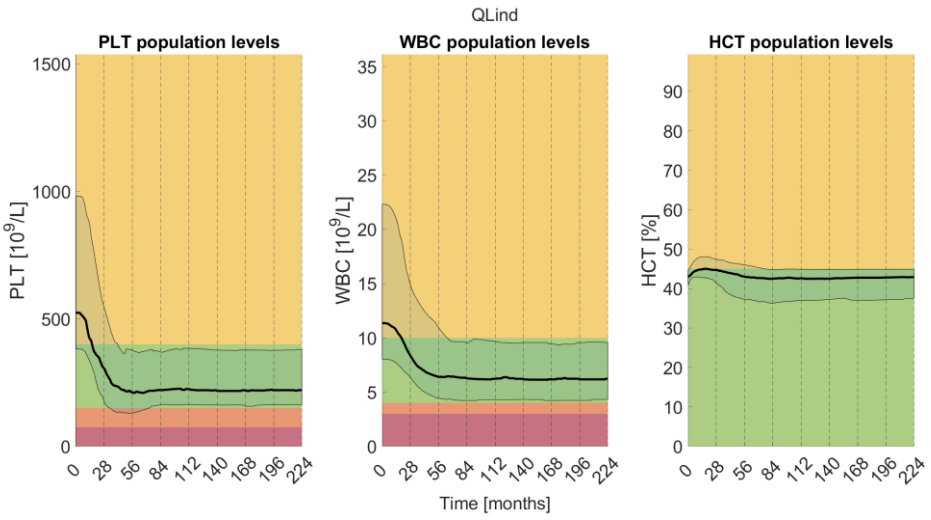


Figure 58: Response rate of PLT, WBC, HCT and CHR for QLpop, QLind and QLind-bay agents in the 8 cycles of treatment.



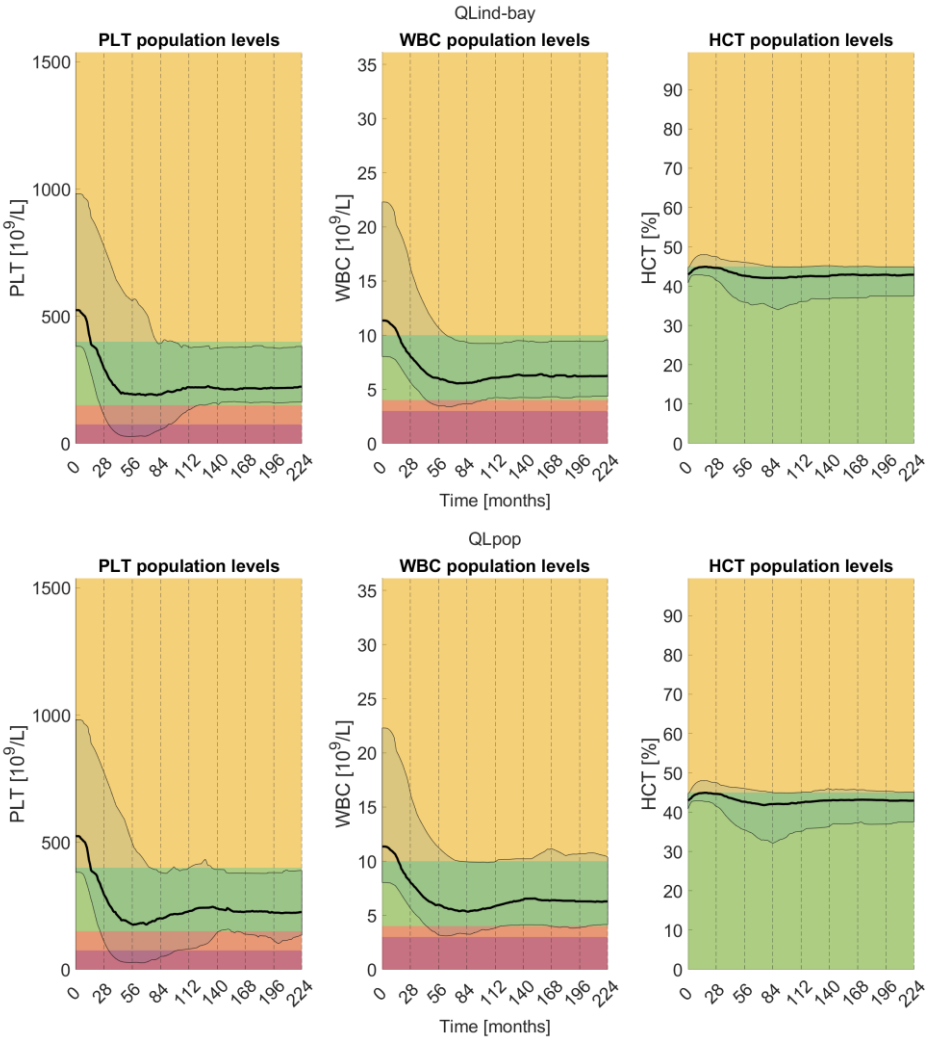


Figure 59: Comparison among QLind, QLind-bay and QLpop agents on the same 112-pateints virtual population. Results are summarized in terms of median (black line) and 90% C.I. of individual profiles within the population (blue shaded areas). Yellow, green, orange and red shaded areas, represents inefficacy, efficacy, moderate and severe toxicity ranges of each haematological parameter according to givinostat clinical protocol described in section 4.1.1.

In particular, as all QLind-bay agents are initialized to QLpop-agent for the first dose selection, their performances are equal after the first cycle (month 1). Differently from QLind-agents, QLpop and consequently also QLind-bay agents use a general set of dosing rules for the first administration thus provoking a higher number of toxicities (i.e., severe thrombocytopenia and/or neutropenia) as highlighted in Figure 59. In the subsequent treatment cycles, QLind-bay agents achieve higher response than those of QLpop-agent as their dose adjustments become more customized due to the paired update of both patient PK-PD parameters though Bayesian MAP and QL-based personal protocols. Consequently, the gap between QLind-bay agents and QLind agents, which hold the top position at each cycle, narrows and,

from the 4th cycle, their performances become very close. At the end of the 8th treatment cycle, QLind and QLind-bay reached very similar CHR rates (97.2% vs 93.4%) and outperformed the 74.7% of QLpop-based protocol.

The Bayesian MIRL strongly reduced the gap with the ideal scenario of QLind-agents starting from the 4th treatment cycle, when at least 4 monitored data points and the baseline values of PLT, WBC and HCT were available for each patient. To investigate the impact of MAP estimations on this result, the following retrospective analysis was performed. Given the dosing schedule administered in the $i-th$ patient by the Bayesian MIRL approach, D_i, its effects were simulated with the real patient parameters, $\boldsymbol{\theta_i}$, and with all the eight sets of individual parameters estimated with the Bayesian MAP at each treatment cycle j, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j}}$. For each PLT, WBC and, HCT profiles simulated with $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j}}$, $f\left(t, \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j}}, D_i\right)$, the root mean squared error ($RMSE_{i,j}$, Eq. 51) with respect to the values obtained using $\boldsymbol{\theta_i}$, $f(t, \boldsymbol{\theta_i}, D_i)$, was computed.

$$RMSE_{i,j} = \sqrt[2]{\frac{1}{T} \cdot \sum_{t}^{T} \left(f\left(t, \widehat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{\boldsymbol{j}}, D_i\right) - f(t, \boldsymbol{\theta_i}, D_i) \right)^2}$$

(51)

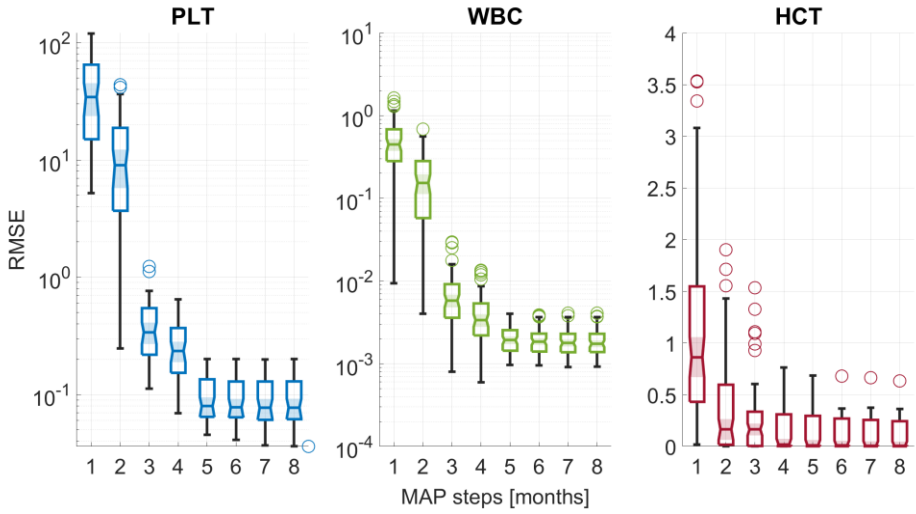


Figure 60: Distribution of the RMSE obtained by simulating givinostat PK-PD model with each cycle-specific estimated parameter set with respect to profile computed with patient original parameters.

Figure 60 reports the distribution of the patients RMSE stratified by treatment cycle for PLT, WBC and HCT. This result shows that from the 4th treatment cycle, the available information was sufficient to obtain good predictions of patients PK-PD response as RMSE become very small for all the three hematological parameters. A further demonstration that four

months are sufficient for the convergence of QLind-bay agents dosing policies to those of QLind-agents is provided by the distributions of the rewards collected by these individual agents at each treatment cycle (Figure 61). At the beginning, as QLind-bay agents were fixed to QLpop agent, their reward distributions are identical. Then, by integrating the information of patient monitoring data in QLind-bay, its dose selections are better remunerated. Starting from the 5th month, their distributions are almost identical. This means that QLind-bay adaptive dosing protocols are very similar to those of QLind-agents.

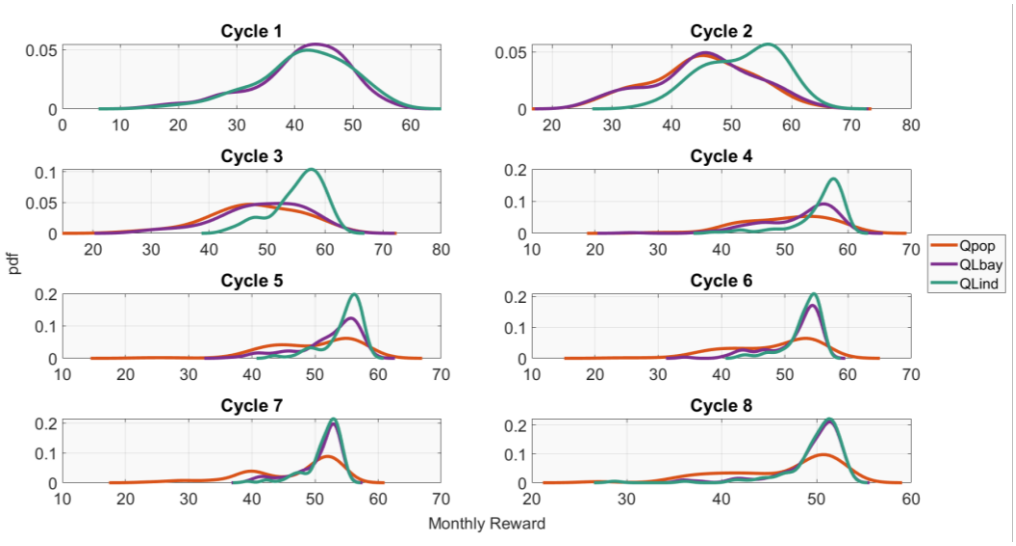


Figure 61: Comparison among the pdf of the reward collected by QLpop, QLind-bay and QLind at each treatment cycle.

6.2. Extending the RL/PK-PD framework to a stochastic treatment response

The second assumption characterizing the MIRL applications in Chapters 3-5 is that the PK-PD model perfectly described the pharmacological response and, consequently, RUV was not considered in the simulations. Neglection of RUV is a simplification and it was demonstrated that the effectiveness of RL algorithms potentially decreases when it is introduced [73]. Consequently, RUV must be adequately managed, with solutions that must depend on what RUV represents, e.g., measurement errors or model misspecification. In the first case, replacing the MDP with a Partially Observable MDP (POMDP) [109] can be a possible solution to formally account for the uncertainty in system state knowledge caused by measurement error. Conversely, this section presents an extension of the MIRL framework, based on QL algorithm (sections 2.1.3 and 2.2.2), suitable in presence of model misspecification.

In a PK-PD modelling framework, the presence of model misspecification is accounted for by defining the observations, z, as a function of the model prediction $f(t, \boldsymbol{\theta_i}, \boldsymbol{x_i}, d)$, where t is the time, $\boldsymbol{\theta_i}$ patient parameters, $\boldsymbol{x_i}$ the individual covariates and d the administered dose, and of a stochastic variable, ϵ. Several RUV models can be adopted and their description can be found in [56–59]. Eq. 52 reports one of the most common RUV model which consists of adding a normally distributed random term proportional to model prediction.

$$z = f(t, \boldsymbol{\theta_i}, \boldsymbol{x_i}, d) \cdot (1 + \epsilon) \text{ , with } \epsilon \sim N(0, \sigma^2). \quad (52)$$

Eq. 52 states that, in presence of model misspecification, the administration of a drug dose d can result in different observations z, which are distributed as a $N(f(t, \boldsymbol{\theta_i}, \boldsymbol{x_i}, d), \sigma^2 f(t, \boldsymbol{\theta_i}, \boldsymbol{x_i}, d)^2)$, whose coefficient of variation (CV) is σ.

In the MIRL based on QL, the finite set of system/patient states, $S = \{s_1, \dots, s_N\}$, is a discretization of the continuous values of z. Each s_i corresponds to a range in which multiple values of z fall within. As seen in the previous MIRL applications, the reward function, ρ, typically depends on the continuous value of the patient state ($\rho(z)$) rather than on its discretization. Consequently, when an action a in $A = \{a_1, \dots, a_M\}$ is performed (i.e., a dose level is selected among, $d_1, .., d_M$), different possible next states, s', can occur due to the stochasticity of z. This also implies that, for the same next state s', different rewards, $r = \rho(z)$, can be observed.

Therefore the conditional probability of observing a next-state, s', and a reward, r, given the previous state-action couple, (s, a), $p(s', r|s, a) = P\{S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a\}$ must be considered (Eq. 2).

QL algorithm leverages a *sample update strategy* during the training phase (Eq. 14) which is based on updating the score of a state-action couple, $Q(s, a)$, with the observed transition $< s', r, s, a >$ (Algorithm 1). However, $< s', r, s, a >$ is only a sample extracted from $p(s', r|s, a)$ and refers to only one possible next-state-reward couple originating from the s, a couple. Moreover, in the sample update strategy, given a state-action couple, all possible next-states and rewards have the same weight, including those less probable. Consequently, this strategy can lead to a biased estimation of the optimal dosing policy [76].

As suggested in [76], in presence of a stochastic system, a more robust update for $Q(s, a)$ is obtained by computing the expected value over all the possible next states and rewards that can be obtained when action a is taken in state s (Eq. 53). This strategy is known as *expected* or *exact* update [76] and, since exact computations are leveraged, learning rate α is not used. Thus, although performing exact updates might be computationally more expensive as all possible next states and rewards should be computed, the number of hyperparameters is reduced with respect to QL.

$$Q(s,a) \leftarrow \sum_{s',r} p(s',r|s,a) \cdot \left[r + \gamma \cdot \max_{a'} Q(s',a')\right].$$

(53)

By using the probability properties, Eq. 53 can be written as in Eq. 54,

$$Q(s,a) \leftarrow \sum_{s',r} p(r|s',a,s) \cdot p(s'|a,s) \cdot \left[r + \gamma \cdot \max_{a'} Q(s',a')\right].$$

(54)

The general approach in Eqs. 53 and 54 can be adapted to the stochastic PK-PD framework due to model misspecifications by considering that rewards depend by z, $r = \rho(z)$, and each $z \in \mathbb{R}$ belongs to a discrete interval $s_i \in S$. Therefore, the exact update formula can be expressed as reported in Eq. 55.

$$Q(s,a) \leftarrow \sum_{s',z \in s'} p(\rho(z)|s',a,s) \cdot p(s'|a,s) \cdot \left[\rho(z) + \gamma \cdot \max_{a'} Q(s',a')\right].$$

(55)

During the training of the QL algorithm, $p(\rho(z)|s',a,s)$ and $p(s'|a,s)$ can be easily computed by assuming a full knowledge of RUV model as well as both patient covariates, $\boldsymbol{x_i}$, and PK-PD parameters, $\boldsymbol{\theta_i}$. Indeed, given the current action (i.e., dose level d), the time of the next monitoring step, t', and the RUV model (e.g., Eq. 52), the distribution of $z \sim N(f(t', \boldsymbol{\theta_i}, \boldsymbol{x_i}, d), \sigma^2 \cdot f(t', \boldsymbol{\theta_i}, \boldsymbol{x_i}, d)^2)$ can easily characterized. Since all values of z can be mapped to an element in S, the $p(s'|a,s)$ can be computed for each possible $s' \in S$. By leveraging PK-PD Monte Carlo simulations, the samples of z can be extracted to characterize the distribution of the reward by applying $\rho(z)$. All these reward values can be stratified following the discrete state in which each z falls, thus $p(\rho(z)|s',a,s)$ is obtained.

When Eq. 55 is used to update $Q(s,a)$, its robustness depends on the correctness of $\max_{a'} Q(s',a')$ [76]. To face system stochasticity and improve $Q(s,a)$ estimation with respect to $\max_{a'} Q(s',a')$, several extensions of QL have been proposed, including Prioritized Sweeping QL which introduces a priority mechanism to update more frequently s,a pairs very impactful on the optimal policy learning (details in section E.5 of appendix E).

In this explorative analysis, a modified version of Prioritized Sweeping QL was combined with the exact update strategy (EQL) of Eq. 55 to address a MIRL precision dosing problem in presence of model misspecification modeled as RUV. The proposed EQL algorithm (section 6.2.1) was directly evaluated on the precision dosing problem of continuous vancomycin infusion in ICU patients (sections 6.2.2-3).

6.2.1. EQL algorithm

The Prioritized Sweeping QL algorithm presented in Algorithm E.1 of Appendix E, was modified by integrating the exact update of Eq.55 (EQL) as reported in Algorithm 4. Its formalization was defined directly considering its integration with PK-PD models and Monte Carlo simulations necessary to characterize $p(\rho(z)|s', a, s)$. By performing exact updates, it is not necessary to define a learning rate, α, as in classic QL. However, two memories, E and D, are introduced to selectively store patient states s^j to re-update and the encountered transitions, respectively. In particular, E contains all s^j for which $\max_{a'} Q(s^j, \mathrm{a})$ has changed following the application of Eq. 55. In EQL algorithm, transitions stored in D are tuples composed by five elements $< S_t, A_t, S_{t+1}^k, \mu_{t+1}^k, \varphi_{t+1}^k >$. Current state-action couple (S_t, A_t) and the $k-th$ possible next state (S_{t+1}^k) are stored analogously to the transitions of other RL algorithms (section 2.1.3 and Appendix A). The two additional saved elements, φ_{t+1}^k and μ_{t+1}^k, are related to S_{t+1}^k and contain the information necessary to re-update a given state. Their definition is obtained from Eq. 55 by considering, for each next state ${s'}_k$, only the summation over $z \in {s_k}'$ (Eq. 56).

$$\sum_{z \in s'_k} p(\rho(z)|s'_k, a, s) \cdot p(s'_k|a, s) \cdot \left[\rho(z) + \gamma \cdot \max_{\mathrm{a}'} Q(s'_k, a')\right] =$$

$$= \sum_{z \in s'_k} p(\rho(z)|s'_k, a, s) \cdot p(s'_k|a, s) \cdot \rho(\mathrm{z}) +$$

$$+ \sum_{z \in s'_k} p(\rho(z)|s'_k, a, s) \cdot p(s'_k|a, s) \cdot \gamma \cdot \max_{\mathrm{a}'} Q(s'_k, a') =$$

$$= \mu_{t+1}^k + \varphi_{t+1}^k \cdot \max_{\mathrm{a}'} Q(s'_k, a').$$

(56)

Algorithm 4: pseudocode of EQL algorithm.

Given: set of N states, set of M actions, discount factor γ, a probability ϵ, a maximum number of training iterations I, selective memory E, transition memory D, patient with a set of covariates $\boldsymbol{x_i}$ and of PK-PD parameters $\boldsymbol{\theta_i}$, PK-PD model $f(t, \boldsymbol{X}, \boldsymbol{\theta})$, probability ϵ

init Q matrix arbitrarily, empty E
loop for each episode (I times):
    Set the current system state to S_0
    loop for each decisional time step t:
        p ← uniform random number ∈ [0,1]

if p<ϵ
    Select action A_t randomly
else
    $A_t \leftarrow arg\ max_a\ Q(S_t, a)$
Perform A_t on the system (dose d_t)
Monte Carlo simulation with $f(t, \boldsymbol{x_i}, \boldsymbol{\theta_i}, d_t)$
loop for each of the possible next states S_{t+1}^k:
    store if absent $< S_t, A_t, S_{t+1}^k, \mu_{t+1}^k, \varphi_{t+1}^k >$ in D
$Q^*(S_t) = \max_a Q(S_t, a)$
update $Q(S_t, A_t)$ with Eq.55
if $Q(S_t, A_t) > Q^*(S_t, a)$
    store S_t in E
loop for each S_t^j in E (J times):
    loop for each S_t^h, A_t^h leading to S_t^j contained in D(H times):
        remove S_t^j form E
        $Q^*(S_t^h) = \max_a Q(S_t^h, a)$
        compute U_j the new contribute of S_t^j on S_t^h, A_t^h with Eq.56
        search in D other possible next states of S_t^h, A_t^h, S_{t+1}^w, with $S_{t+1}^w \neq S_t^j$
        Compute U_w for each S_{t+1}^w
        update $Q(S_t^h, A_t^h)$ with $U_j + \sum_w U_w$
        if $Q(S_t^h, A_t^h) > Q^*(S_t^h)$:
            store S_t^h in E
randomly set current system state to S_{t+1} according to $p(S_{t+1}|S_t, A_t)$

6.2.2. Continuous vancomycin infusion regime in ICU patients

Vancomycin is a glycopeptide antibiotic used to treat and prevent various bacterial infections caused by gram-positive bacteria, including methicillin-resistant *Staphylococcus aureus* (MRSA) [132]. This drug can be given either orally or intravenously, and, for both administration routes, an adaptive dosing protocol based on the monitoring of plasma concentrations is followed [34,133]. Indeed, vancomycin dosing is challenging due to its large IIV in PK response, the narrow therapeutic window and the severe AEs induced by excessive exposures such as nephrotoxicity, hypotension, and hypersensitivity reactions [134]. Since critically ill patients require stable exposure, higher target attainment and lower rate of nephrotoxicity [135], Vancomycin is administered through a continuous 24-hours infusion regimen adjusted based on daily monitoring of PK concentrations. The target concentration range is 15-25 mg/L; values <15 mg/L are considered

ineffective while concentrations >25 mg/L can lead to moderate (25-30 mg/L) and severe toxicities (>30 mg/L). Vancomycin precision dosing in ICU patients can be subdivided into three stages. First, a loading dose based on patient weight with a rate of 10 mg/minute is applied. Then, the initial vancomycin dose to be infused on 24 hours is selected following the evaluation of patient renal functions. Finally, the maintenance stage is based on the adaptive dosing rules summarized in Table 13. In the protocol reported in [136], the available doses of vancomycin are {250, 500, 1000, 1500, 2000, 2500, 3000, 3500} mg and PK monitoring starts after the initial 24-hours vancomycin infusion.

Table 13: Clinical indications for adjusting daily vancomycin doses.

Vancomycin concentration	Dose adjustment
<15 mg/L	Increase the dose of 500 mg.
[15-25] mg/L	Maintain current dose level.
(25-30] mg/L	Decrease the dose of 500 mg. If the current dose is 500 mg, decrease to 250 mg.
>30 mg/L	Stop the infusion for 6 hours and then restart with a lower dose level (dose decrease is arbitrarily established by clinicians at ward round without any restriction).

6.2.3. Formalization of vancomycin precision dosing problem within MDP framework

Before applying the implemented EQL, the precision dosing problem relating to the vancomycin case study was formalized as a MDP (section 2.1.1). Therefore, the components of MDP i.e., system states, agent actions and reward function, were defined based on the clinical scenario described on section 6.2.2. However, to test the proposed EQL algorithm in a more challenging scenario, it was hypothesized that vancomycin PK monitoring begins immediately after the administration of the loading dose, rather than collecting the first sample after the initial 24-hours infusion.

Both state and action sets were defined in a discrete fashion, according to EQL framework. System/patient state are based on serum vancomycin concentration levels which are periodically monitored to guide dose adjustments. Reward function was designed to define patient specific adaptive dosing protocol bringing and maintaining vancomycin serum concentration levels within the target range of [15-25] mg/L. The available actions to EQL-agent were defined in accordance with the vancomycin clinical protocol documented in [136] and summarized in section 6.2.2.

A literature Pop-PK model of vancomycin developed on ICU patients receiving a continuous infusion treatment regimen was integrated within

EQL framework to simulate patient response to different dosing schedules (further details in section E.6 of Appendix E). Differently from the original RUV model (EQ. E.6), a simpler proportional RUV model (Eq. 52) was adopted to better understand the obtained results.

Furthermore, a 15-days treatment duration was assumed for the analysis based on previous findings from studies on continuous vancomycin treatment in ICU patients [137,138].

From an implementation perspective, in this investigation both vancomycin Pop-PK model and EQL algorithm were coded in MATLAB [115].

6.2.3.1. Reward function

The reward function was defined according to the daily monitored vancomycin serum concentration level, $[Vanco]_{daily}$, which is a stochastic variable due to RUV. As reported in Eq. 57 and illustrated in Figure 62, the function gives the highest remuneration to dosing strategies leading to the middle of the efficacy range (i.e., [17.5,22.5] mg/L interval). Then, reward values linearly decrease as $[Vanco]_{daily}$ moves away from the [17.5,22.5] mg/L range, though they remain within the efficacy concentration window. Conversely, an exponential decay of the reward values occurs if $[Vanco]_{daily} < 15\ mg/L$ (inefficacy) or $[Vanco]_{daily} > 25\ mg/L$ (threshold of moderate toxicity). Values returned by the reward function become close to 0 in presence of very low effective concentrations ($[Vanco]_{daily} < 10\ mg/L$) or severe toxicities ($[Vanco]_{daily} > 30\ mg/L$).

$$Reward\left([Vanco]_{daily}\right) = \begin{cases} 4 \cdot \exp\left(\alpha \cdot \left|15 - [Vanco_{daily}]\right|\right) & if\ [Vanco_{daily}] < 15\ mg/L \\ 1.6 \cdot [Vanco_{daily}] - 18 & if\ [Vanco_{daily}] \in [15,17.5)\ mg/L \\ 10 & if\ [Vanco_{daily}] \in [17.5, 22.5] mg/L \\ -1.6 \cdot [Vanco_{daily}] - 46 & if\ [Vanco_{daily}] \in (22.5, 25] mg/L \\ 4 \cdot \exp\left(\alpha \cdot \left([Vanco_{daily}] - 25\right)\right) & if\ [Vanco_{daily}] > 25\ mg/L \\ \alpha = 0.2 \cdot \ln 0.025 \end{cases}$$

(57)

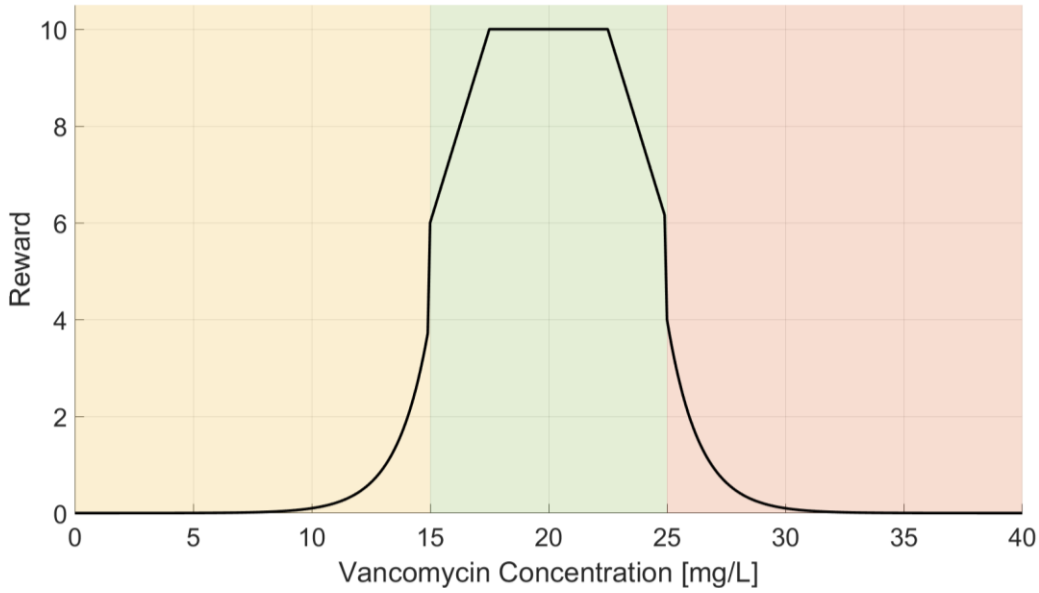


Figure 62: Reward function for vancomycin precision dosing problem based on its serum concentration levels. Yellow shaded area represents inefficacy range, the green and the red ones, efficacy and toxicity (both moderate and severe), respectively.

6.2.3.2. System/Patient states

Patient health status was described by a tuple of three elements, $X = \{Vanco_{Discr}, PrevDose, Stop_{flag}\}$. The first, $Vanco_{Discr}$, accounts for the $[Vanco]_{daily}$ observation which is discretized in the eight levels, as detailed in Eq. 58. In particular, the ranges are defined according to the subdivision proposed by the clinical protocol (Table 13), and a $+/-$ sign is added to distinguish between the monitoring after the loading dose (i.e., selection of the first vancomycin infusion over 24 hours) and the maintenance stage.

$$Vanco_{Discr} = \begin{cases} -1 & [Vanco]_{daily} < 15\ mg/L\ (post-loading\ dose\ monitoring) \\ -2 & [Vanco]_{daily} \in [15{,}25] mg/L\ \ (post-loading\ dose\ monitoring) \\ -3 & [Vanco]_{daily} \in (25{,}30] mg/L\ \ (post-loading\ dose\ monitoring) \\ -4 & [Vanco]_{daily} > 30\ mg/L\ (post-loading\ dose\ monitoring) \\ 1 & [Vanco]_{daily} < 15\ mg/L\ (maintenance\ stage) \\ 2 & [Vanco]_{daily} \in [15{,}25] mg/L\ (maintenance\ stage) \\ 3 & [Vanco]_{daily} \in (25{,}30] mg/L\ (maintenance\ stage) \\ 4 & [Vanco]_{daily} > 30\ mg/L\ (maintenance\ stage) \end{cases}$$

(58)

Differently, $PrevDose$ assumes values in {250, 500, 1000, 1500, 2000, 2500, 3000, 3500} mg and represents the previous administered dose of vancomycin. Its definition was made according to the clinical protocol described in section 6.2.2 [136]. Finally, $Stop_{flag}$ is a flag which is 1

whether treatment was temporary stopped for 6 hours, 0 otherwise. All negative values of $Vanco_{Discr}$ were combined with all elements in $PrevDose$, obtaining an initial set of 32 states. Then, the other states were obtained by considering all possible combinations between $Vanco_{Discr}$>0, $PrevDose$ and $Stop_{flag}$ (n=64). Finally, a state coding for treatment beginning, S_0, was added. Overall, a 97-dimensional states space was obtained.

6.2.3.3. EQL Agent actions

Actions of EQL agent were designed considering the clinical available vancomycin dose levels reported in [136] (section 6.2.2). Table 14 summarizes the constraints in dose selection that could be explored by the algorithm during the training stage. EQL agent could select both the loading dose, and the initial 24-hours infused amount from all the available vancomycin dosages, {250, 500, 1000, 1500, 2000, 2500, 3000, 3500} mg. For the loading dose, the infusion rate of 10 mg/minutes was maintained accordingly to the original protocol. Then, at each decisional step of the maintenance stage, EQL could choose between maintaining the current dose level ($D=$) or performing stepwise changes (i.e. increase/decrease by one level, $D+/D-$), independently form the currently observed vancomycin concentration.

Differently from the case studies in Chapters 3-5, strict safety constraints were not imposed in presence of moderate/severe toxicities. Indeed, the EQL agent was allowed to decide whether to temporarily interrupt (6 hours of stop) treatment and then resume at a lower dose level, or not. The resumption dose was constrained to a drug level lower than that administered prior to the interruption.

The choice of performing a temporary interruption was allowed even when vancomycin concentrations were within the effective range, but not when concentrations were too low and ineffective.

Table 14: Summary of the actions available to EQL agent for each patient state.

Patient state	**EQL Agent available actions**
Loading dose selection i.e., initial state S_0	250, 500, 1000, 1500, 2000, 2500, 3000, 3500 mg with a 10 min/hours rate of infusion
Initial 24-hours vancomycin infusion, i.e., for each state with $Vanco_{Discr} < -1$	250, 500, 1000, 1500, 2000, 2500, 3000, 3500 mg
In presence of treatment inefficacy during the maintenance stage, i.e., for each state with $Vanco_{Discr} = 1$	D+, D-, D=

In presence of efficacy, moderate/severe toxicities during the maintenance stage	D+, D-, D=, stop treatment for 6 hours and resume for 18 hours with a dose level in {250, 500, 1000, 1500, 2000, 2500, 3000, 3500} mg $<PrevDose$

6.2.3.4. Evaluation Framework of EQL agent

The performances of the EQL algorithm were evaluated on a typical patient whose vancomycin PK model parameters were fixed to the population ones (θ_{CL} and θ_{Vol} reported in Table E.7). To this end, the EQL-agent was tested with three different scenarios of stochasticity, corresponding to three different CVs for the proportional RUV model. Low (CV=10%), moderate (CV=20%) and high (CV=30%) RUV impact on vancomycin PK were used in the analysis. To account for the stochasticity of the process, the 15-days treatment scenario on the typical patient was repeated 10000 times and the distribution of the cumulative discounted rewards computed for each replica was considered in as evaluation metric.

Furthermore, to obtain a robust assessment, EQL-agent was benchmarked by a i) QL-agent, named QLc-agent, which is trained by considering RUV with classic QL algorithm (section 2.1.3) and ii) a QL-agent, named QLdet-agent, trained neglecting RUV. The comparison between EQL-agent and QLc-agent allows the evaluation of the algorithm robustness when RUV is properly handled with respect to the standard method. Differently, the comparison between EQL-agent and QLdet-agent, which represents the MIRL framework adopted in Chapters 3-5 where RUV was neglected, allows to assess the drawbacks of ignoring the stochasticity due to model misspecifications. Finally, by comparing QLdet and QLc, it is possible to assess the effect of neglecting RUV in QL training in presence of a wrong modelling framework.

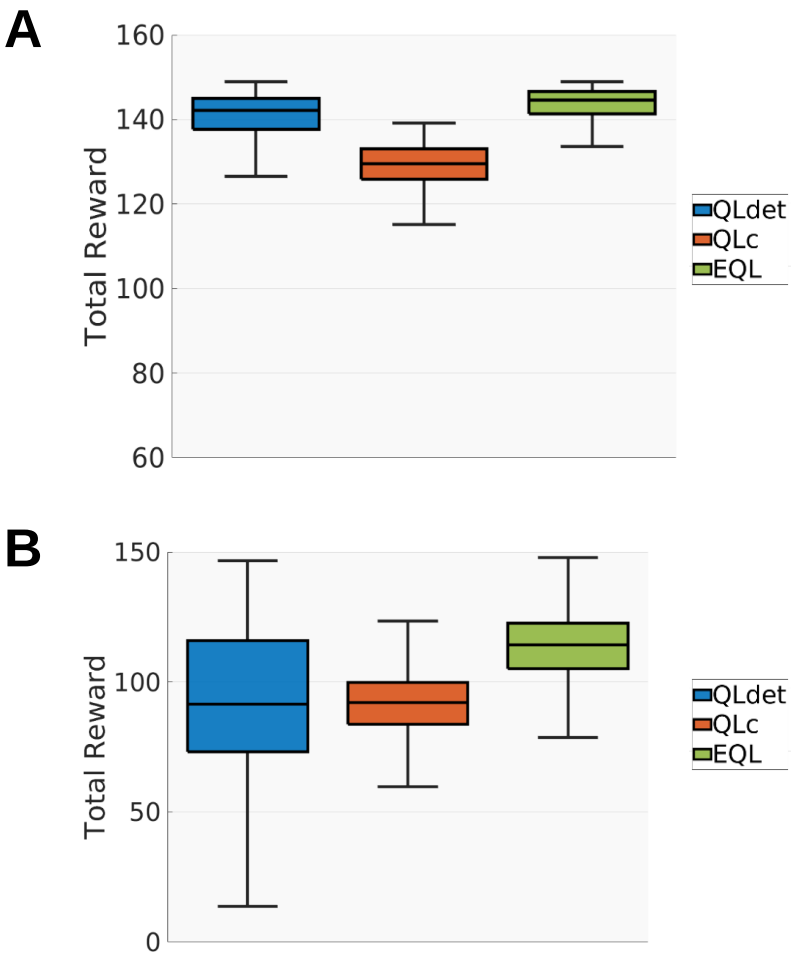
A further analysis was conducted to evaluate the impact of a misspecification in the RUV model on EQL-agent performances. To this end, the EQL-agent was trained using a CV in {10%, 20%, 30%} for the proportional RUV model (Eq. 52), then the remaining two were used as real proportional RUV to simulate the effects of agent optimal dosing policy on the typical patient. This strategy allows to assess the effect of both overestimating and underestimating the RUV CV. As reported in Table 15, for each of these assessments, the adopted benchmark was the scenario in which training and test shared the same CV of RUV model (i.e., perfect estimation of the CV).

Table 15: Summary of the scenarios considered to assess the effect of misspecifications of RUV model on EQL performances.

RUV CV of EQL Training	RUV CV of EQL Test	
	Underestimation/Overestimation	Benchmark
10%	-/20%,30%	10%
20%	10%/30%	20%
30%	10%, 20% / -	30%

6.2.3.5. Results

EQL, QLc, and QLdet agents were compared in customizing the treatment for the typical patient by leveraging the distribution of the total discounted rewards obtained for each replica of the treatment scenario. Results are presented in Figure 63 stratified for each RUV CV scenario. Algorithm hyperparameters are reported in section E.7 of Appendix E.



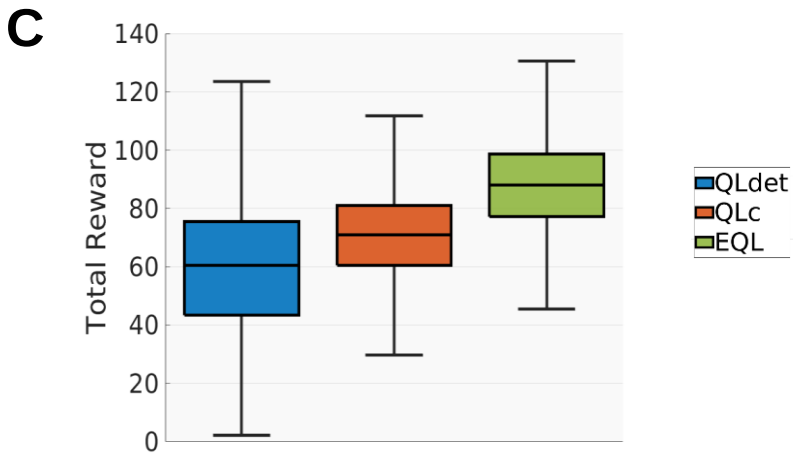


Figure 63: Comparison between QLdet, QLc and EQL agents in the personalization of the typical patient by considering the sum of the discounted reward collected within 15 days of treatment. Each panel corresponds to a RUV CV scenario (A=10%, B=20%, C=30%).

In particular, in each RUV scenario the EQL-agent obtained higher cumulative discounted rewards with lower variance than the QLc and QLdet agents. This result confirms that the novel introduced approach is more robust in presence of stochasticity due to model misspecifications.

In case of high CVs (20 and 30%, panels B and C of Figure 63), classical QL performed better when RUV is considered during the training, i.e., QLc, than when RUV is neglected, i.e., QLdet. Indeed, neglecting the stochasticity in patient response (i.e., RUV) avoids the QLdet to consider some state-action couples thus provoking a lack of dose suggestions for some patient conditions. In the simulation study, in these cases, dose was randomly selected.

QLc advantage over QLdet is not consistently observed across all cases. Interestingly, in presence of low RUV (CV=10%, panel A of Figure 63) neglecting the stochasticity due to model misspecifications does not dramatically lowers QLdet performances as it reached a distribution of the total discounted rewards close to that of EQL-agent and higher than QLc-agent. The comparison of dose suggestions in the three algorithms with RUV CV=10% (Figure 64) explains why QLdet and EQL agents perform better than QLc-base one. Indeed, it shows that QLc-agent generally adopts lower dose levels than QLdet and EQL ones. Consequently, this lower exposure is not effective for bringing vancomycin concentration in the target range and maximizing the total discounted reward. This behaviour of the QLc-agent is provoked by the bias due to the sampling update. Indeed, although the observation of a toxicity provoked by high doses is rare, it was encountered during the training and the low reward obtained prevent their selection. This more conservative approach of QLc-agent occurred also with higher rewards (Figure 65 and Figure 66) but it is less penalizing in presence of higher RUV CVs.

These results highlight that considering RUV in MIRL framework is important to avoid suboptimal dosing policies, especially when its impact is moderate or high not. At the same time, it should be properly handled, otherwise suboptimal policies can be still to estimated, with worse performances than the case in which it is neglected.

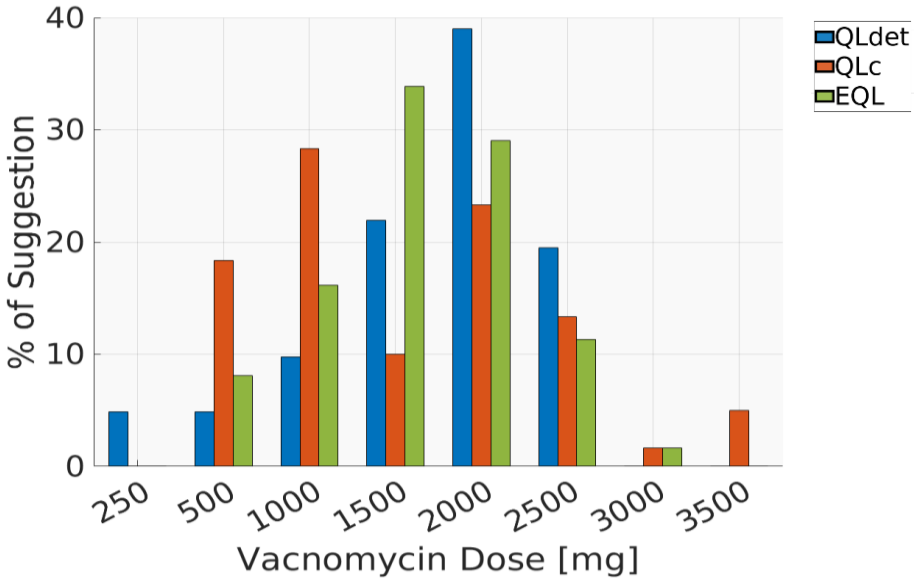


Figure 64: Distribution of dose suggestions for QLdet, QLc and EQL agents with a RUV CV of 10%.

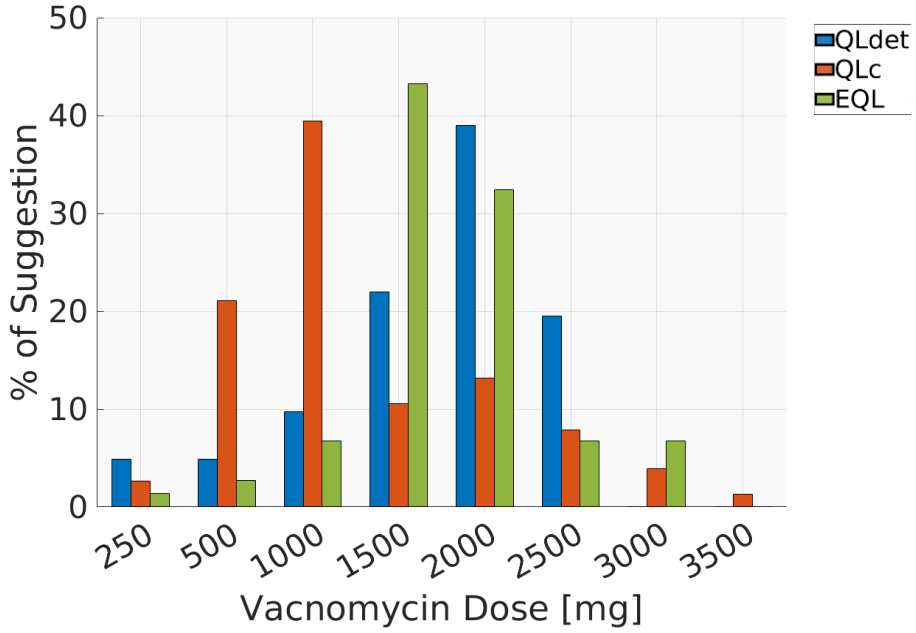


Figure 65: Distribution of dose suggestions for QLdet, QLc and EQL agents with a RUV CV of 20%.

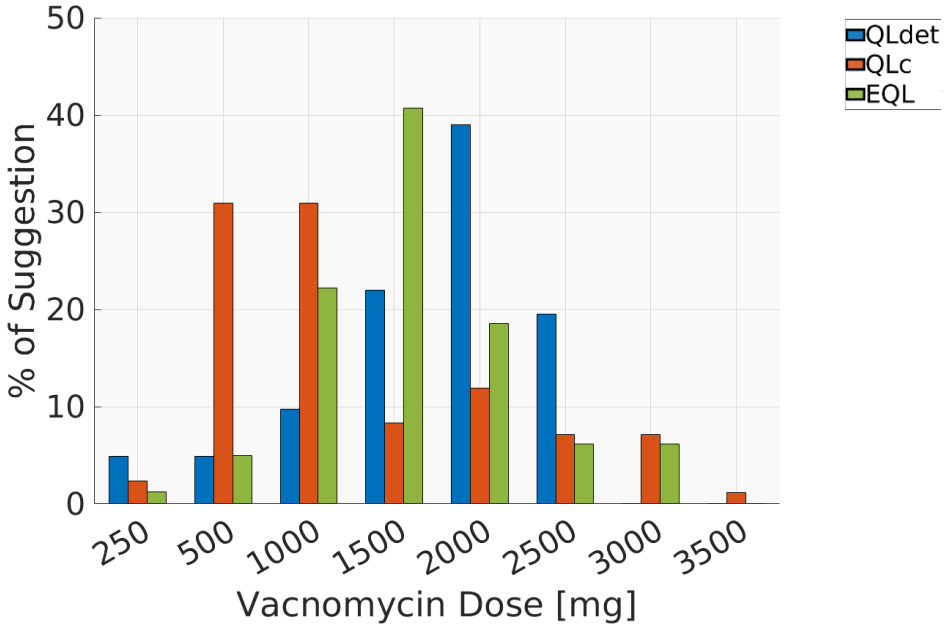
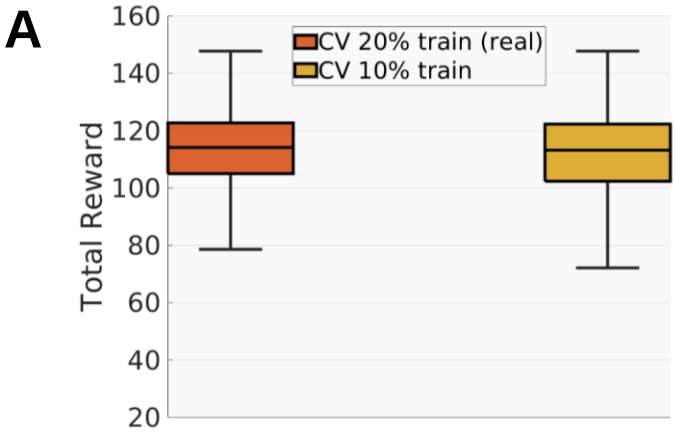


Figure 66: Distribution of dose suggestions for QLdet, QLc and EQL agents with a RUV CV of 30%.

To assess the robustness of EQL-agent in presence of wrong RUV quantification with the proportional error model (Eq. 52), the simulation scenarios on the typical patient summarized in Table 15 were considered. Figure 67 shows the obtained results by stratifying the distributions of the total discounted rewards for underestimation (panels A and B of CV=20% and CV=30%, respectively) or overestimation cases (panels C and D of CV=10% and CV=20%, respectively).



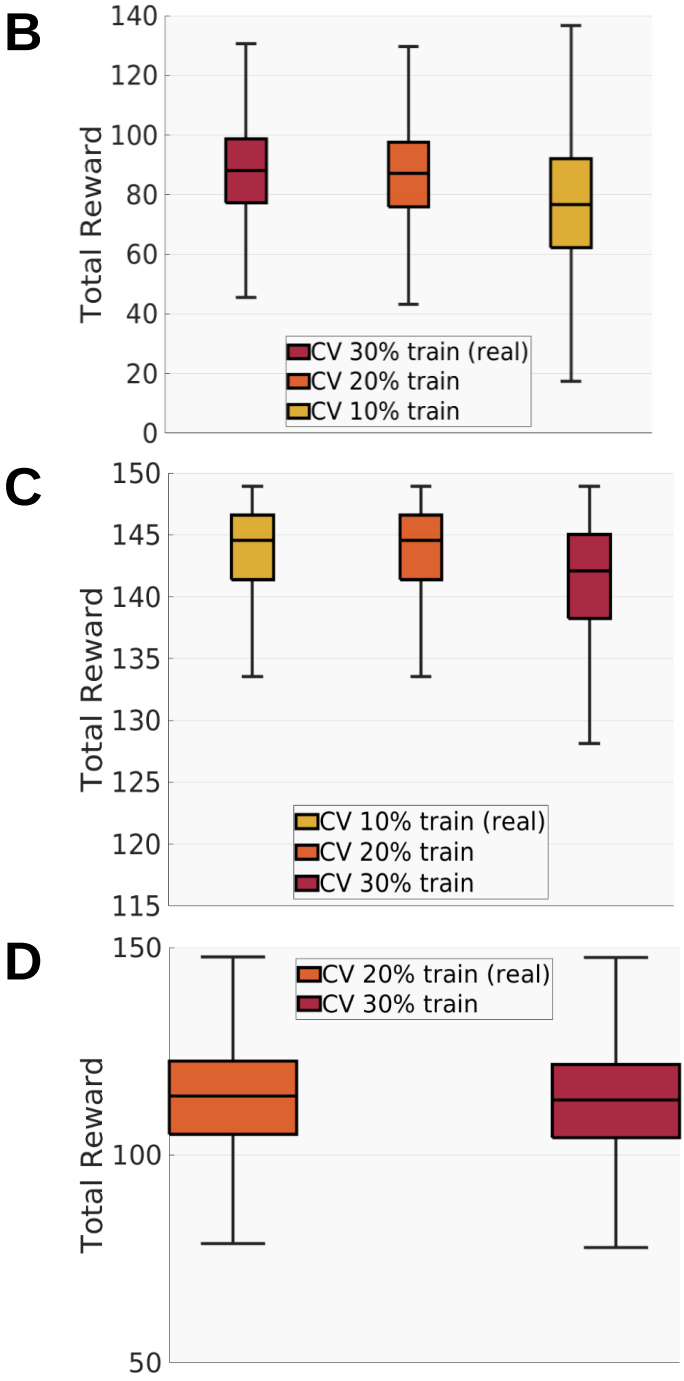


Figure 67: Effects of a wrong RUV CV estimation of the total discounted rewards. Panels A and B show the scenarios of underestimating CVs of 20% and 30%, respectively. Panels C and D are focused on the overestimation of CVs of 10% and 20%, respectively.

In particular, this analysis shows that extreme overestimations (actual CV=10%, estimated CV=30%) or underestimations (actual CV=30%, estimated CV=10%) can dramatically alter EQL-agents performances by decreasing the median sum of total rewards and increasing their variance.

Differently moderate underestimations/overestimations preserve the EQL-agent performances.

6.3. Discussions

In this Chapter, two extensions of the proposed individual-oriented MIRL approach presented in section 2.2.2 were developed to address the limitations of its originally formalization, which actually prevent the implementation of the RL/PK-PD paradigm in clinical practice.

First, starting from the seminal works in [63,80], a Bayesian MIRL framework was developed to overcome the assumption made for the case studies in Chapters 3-5 that patient digital twins were well-characterized since the beginning of the treatment (i.e., the individual PK-PD model parameters were known). Indeed, this hypothesis does not always hold, as individual parameter estimation requires the availability of individual data and, consequently, needs to be performed during the monitoring of the treatment.

The introduced Bayesian MIRL leverages RL techniques within the MIPD framework. In this case, the prior knowledge on patient pharmacological response includes individual covariates and a Pop-PK-PD model, as in MIPD, and also a general RL-based dosing protocol already derived on a population of patients. Before treatment begins, individual RL-agents cannot be defined as patient-specific PK-PD model parameters are unknown and, consequently, the personalized RL-based dosing strategies are fixed to the general one to select the initial treatment dose for each patient. Then, at each monitoring step, PK-PD samples collected from the patient are used to update the estimation of both the patient-specific model parameters characterizing the digital twin and the individual RL-agent. The performances of this novel Bayesian MIRL framework were evaluated on a simulation study using a simplified version of the givinostat precision dosing problem presented in section 4.1.1. QL was used as RL algorithm also in this explorative analysis.

In particular, the simpler version of givinostat precision dosing problem consisted in a reduced number of drug dose levels accordingly to early clinical trials [45]. Consequently, as detailed in section E.1 of Appendix E, the formalization of givinostat case study within MDP framework was different from that reported in section 4.1.2. Indeed, less system/patient states and constraints in action selection were adopted in this case.

A heterogeneous 112-patients virtual population generated with the stratified random sampling strategy described in section E.2 of Appendix E was used to evaluate the individual QL-agents of the Bayesian MIRL paradigm (QLind-bay agents). In particular, before treatment begin, all QLind-bay agents were fixed to the general adaptive dosing strategy of a QLpop-agent already trained on another virtual patient population (details in section E.4 of Appendix E). Consequently, the initial treatment dose was

determined for each virtual patient according to general dosing rules based on individual baseline PLT, WBC and HCT levels. Then, at each monitoring step, the currently observed values of the haematological parameters as well as their previous ones, were used to update patient PK-PD model parameters by performing Bayesian MAP estimation. The updated patient digital twin was then used to train the QLind-bay agent thus refining the current patient optimal policy and informing the next-dose selection coherently to the latest individual characterization.

The performances of the QLind-bay agents were benchmarked against both the QLpop-agent and the QLind-agents on the same virtual population. The QLpop-agent (section 2.2.1) represents the case in which a general RL-based protocol optimized for an entire patient population is used to guide patient dosing without performing individually tailored dosing strategies. Differently, the QLind-agents reflect the less realistic RL-based scenario in which it is assumed that patient PK-PD parameter are fully known before the treatment begins.

From an implementation perspective the Bayesian MIRL approach was developed extending the hybrid MATLAB-NOMEM framework presented in section 4.1.2.4. Specifically, MATLAB was still used to code QL algorithm, differently, NONMEM was leveraged to perform both PK-PD simulations and Bayesian MAP. All simulations were made in absence of RUV to have a clearer depiction of how this novel approach perform.

From the comparison among QLind-bay, QLind and QLpop agents, the ideal scenario in which perfect patient knowledge is assumed from treatment begin obviously obtained the best results. However, QLind-bay was able to reach interesting performances very close to those of the QLind-agent. In particular, as QLind-bay agents were initialized to the dosing policy of QLpop-agent for the first dose selection, their performances after the first treatment cycle coincided and were lower than QLind-agent. From the 2nd cycle onward, QLind-bay agents begin to outperform QLpop as their dosing strategies become more personalized through Bayesian updates of patient PK-PD parameters By the 4th treatment cycle, QLind-bay agents nearly match the performance of the ideal QLind-agents and, at the 8th treatment cycle (last monitoring step), both strategies achieved a similar CHR rate (QLind-bay=93.4%, QLind=97.2%, respectively) and outperformed QLpop (74.7%).

The retrospective analysis on Bayesian MAP estimations performed before each QLind-bay update and dose suggestion revealed that from the 4th cycle, sufficient data was available to make accurate predictions of the real patient PK-PD response. Therefore, as patient characterization improves, also QLind-bay adaptive dosing protocols become better over time converging to those of QLind-agents. This result is confirmed by the comparison of rewards collected at each treatment cycle by Qlind-bay, QLind, and QLpop agents. In particular, starting from the 5th cycle, QLind-bay agents dosing strategies receive almost the same rewards of QLind-

agents, indicating that the Bayesian RL/PK-PD framework effectively adapts its dosing strategies over time to approach the ideal scenario.

Overall, the developed Bayesian MIRL framework showed interesting results thus providing a valid solution to the issue of estimating patient PK-PD parameters during treatment monitoring to robustly train personal RL-agents. Despite the good performances, some refinements could be investigated to further improve the robustness and the assessment of this methodology. In the performed explorative analysis, RUV was neglected to improve the clarity of the obtained results, therefore another investigation is required. Moreover, another possible investigation is related to the Bayesian estimation techniques. The implemented Bayesian MIRL approach leveraged Bayesian MAP estimation which returns the most likely patient PK-PD parameters set [139–141]. However, this strategy neglects the uncertainty on the estimated parameters, due to, for example, the sparsity of the monitored PK-PD dynamics. Consequently, the RL-based adaptive dosing rules might be suboptimal in presence of high parameter uncertainty. Therefore, an interesting extension of the introduced Bayesian-MAP-based MIRL is the integration of RL/PK-PD with full Bayesian techniques which estimates the distributions of patients PK-PD parameters rather than a single values and, consequently, to estimate dosing rules accounting for parameters uncertainty [80,142]. Finally, from an implementation perspective, an alternative framework to the MATLAB-NONMEM combination could be explored as it showed high computational times for the RL training, similarly to those discussed in section 4.3. This aspect becomes central when moving toward a real-world application of this methodology. As the re-training of RL-agent occurs at each patient monitoring before the subsequent dose adjustment, the updated RL dosing rules should be available within the time frame spanning from patient observation to clinical decision in order to leverage the RL-based suggestion.

A second novel approach was developed to address the presence of RUV representing model misspecifications. To this end a novel extension of MIRL paradigm exploiting QL was developed to handle the stochasticity of reward function and states transitions. In particular, this method leverages PK-PD models and Monte Carlo simulations to characterize the probabilities of all possible next patient states and rewards following a dose selection and perform an exact updated of the estimated optimal dosing policy.

To allow a robust estimation of the patient optimal dosing strategy in presence of model misspecifications, an extension of Prioritized Sweeping QL based on exact updates of the Q-function, $Q(s,a)$, (EQL algorithm) was developed.

The developed EQL algorithm was then evaluated on the clinically relevant precision dosing problem of vancomycin continuous infusion in ICU patients. This glycopeptide antibiotic is effective against gram-positive bacteria, including MRSA, and is administered in ICU patients through a continuous infusion regimen to obtain stable drug exposure and reduce nephrotoxicity. Due to its narrow therapeutic window, significant IIV in PK

response and the high risk of severe adverse effects in such fragile population, the continuous infusion of vancomycin follows an adaptive dosing protocol. Drug serum concentration is monitored and dose adjusted to maintain Vancomycin levels within the efficacy range of [15,25] mg/L. Vancomycin levels below 15 mg/L result ineffective while concentrations above 25 mg/L are associated with toxicity. Vancomycin precision dosing in ICU patients involves three phases: an initial loading dose based on weight, the selection of the first 24-hour infusion dose considering renal function, and maintenance dosing based on daily monitoring of the serum concentration.

Before applying the EQL algorithm to the vancomycin precision dosing problem, it was first formalized as an MDP. Therefore, system states, agent actions, and reward function were defined accordingly to the clinical scenario [136]. In particular, the reward function was designed to guide dosing adjustments to maintain serum vancomycin levels within the target range of 15-25 mg/L. To create a more challenging test for the EQL algorithm, it was hypothesized that PK monitoring would begin immediately after the loading dose rather than following the initial 24-hour infusion [136].

A literature-based Pop-PK model specific to ICU patients was integrated into the EQL framework to simulate responses to different dosing schedules [137]. The original RUV model was replaced by a simpler proportional RUV model to obtain clearer results. The analysis assumed a 15-days treatment duration, aligning with previous studies on continuous vancomycin treatment in ICU patients [137,138].

The EQL-agent performances were evaluated on a typical patient using fixed population vancomycin PK parameters (values in Table E.7 of Appendix E) testing three levels of stochasticity (low, moderate, and high) corresponding to three different CVs in the proportional RUV model (10%, 20% and 30%). The 15-day treatment was simulated 10,000 times to account for stochasticity, with the cumulative discounted rewards analyzed.

The EQL-agent was compared to two benchmarks: the QLc-agent, which considers RUV in the classic QL algorithm, and the QLdet-agent, which ignores RUV during training. These comparisons were introduced to i) assess the robustness of EQL-agent in handling RUV with respect to standard approach (QLc) ii) quantify the consequences of neglecting RUV due to model misspecifications (QLdet-agent) vs using framework that implicitly manage this stochasticity (EQL-agent) iii) evaluate the impact of neglecting a wrong modelling framework (represented by RUV) in QL training (QLdet vs QLc).

EQL-agent outperformed both QLc and Qldet agents in all RUV CV scenarios by achieving higher total discounted rewards, thus confirming that the novel introduced approach is more robust in presence of stochasticity due to model misspecifications. QLc-based dosing strategies resulted suboptimal in all three cases with respect to EQL-agent. In particular, QLc-agent typically administers lower doses compared to the EQL-agent. As a result, QLc-based dosing strategies provoke lower exposures often failing to

achieve both the target vancomycin concentration range and high sum of the discounted rewards. This conservative dosing behavior in the QLc-agent is due to the sampling bias of the QL algorithm (sections 2.1.3 and 6.2.1) in the update process. Indeed, such algorithm does not give a lower weight to rare high-dose-related toxicities and, consequently, their selection is strongly penalized. This characteristic leads the QLc-agent to obtain lower performances than the QLdet one in presence of the CV=10% scenario. Interestingly, in this scenario Qldet-agent performed similarly to EQL, thus remarking that in presence of low RUV due to model misspecifications, the stochasticity of patient response can be neglected. These results highlight that, considering RUV in MIRL framework is important to avoid suboptimal dosing policies, especially when its impact is moderate or high not. At the same time, it should be properly handled, otherwise suboptimal policies can be still to estimated, with worse performances than the case in which it is neglected

Additionally, the impact of RUV model misspecification on EQL-agent performances was analyzed by training the algorithm with one CV level and testing it against the others, evaluating the effects of overestimating or underestimating RUV. The benchmark for each assessment was the scenario where the training and testing CVs matched perfectly. The obtained results showed that extreme errors in RUV model estimation can significantly impact EQL-agent performance. Specifically, overestimating the real CV of 10% as 30% or underestimating the real CV of 30% as 10% could substantially decrease the median total rewards and increase their variance. In contrast, moderate errors in RUV estimation, whether underestimations or overestimations, had a lower impact and generally preserved EQL-agent performance.

Overall, EQL algorithm performed very well in presence of RUV due to model misspecifications. Further step forward could be expanding this framework to account for other sources of uncertainties, for example those coming from PK-PD parameter estimation during patient monitoring as in the full Bayesian approach.

In conclusion, although it is of interest to investigate some small refinements, both the methodologies presented in this chapter address the current limitations of the individual oriented MIRL paradigm (section 2.2.2) showed promising results. Therefore, the next step is to integrate the EQL within Bayesian MIRL to simultaneously address model misspecifications and the estimation of patient PK-PD parameters during the treatment with their respective uncertainties.

Chapter 7

Overall Conclusions

Precision dosing is currently at the core of the debate within the precision medicine revolution, as this approach can definitively overcome all the limitations of the classical one-size-fits-all paradigm. Indeed, this patient-centric workflow raised the attention of both clinical community and regulatory agencies, and it has been recommended for several classes of compounds showing a narrow therapeutic window, significant IIV, severe or irreversible AEs due to overdosing and/or for diseases with serious consequences of undertreatment. Precision dosing can address all these issues by tailoring drug administration in a specific patient at a given time according to individual factors that influence treatment response. Adaptive dosing strategies are central in this workflow as they perform dose adjustments customized at patient level based on the monitoring of efficacy and/or toxicity treatment biomarkers in the individual.

Pharmacometrics modelling supports precision dosing tasks and the MIPD approach rapidly gained momentum. The recent outbreak of AI/ML in pharmacometrics represents an opportunity to explore novel hybrid methodologies coupling PK-PD modelling with AI/ML to further improve MIPD in the context of adaptive dosing strategies. Among them, RL, a ML subfield encompassing different algorithms to solve sequential decision-making processes, naturally fits to precision dosing problems based on periodic patient monitoring and adaptive dosing strategies. Therefore, there is great interest within the pharmacometrics community in exploring the integration of PK-PD modelling and RL, i.e., the MIRL paradigm, to support precision dosing. Although different works have already been published [30,39,73,79,80,83–86,88,89,89,91,92], several open questions on MIRL approaches are still unsolved [74].

The MIRL approach presented in the literature generally performs well in populations with low to moderate IIV but can be suboptimal when it becomes high. Furthermore, available studies often considered simplified and unrealistic precision dosing problems. They were mostly focused only on a

single efficacy or toxicity endpoint and neglected other relevant biomarkers or constraints in dose selection. Additionally, the effectiveness of MIRL in optimizing long-term outcomes and managing concomitant drugs is still not well understood.

The aim of this thesis was to further investigate the integration of RL with PK-PD models to deliver precision dosing, addressing the limitations of the works currently available in the literature. In particular, following the seminal work of Maier et al. (2021) [80], a novel MIRL framework to learn clinically acceptable adaptive dosing strategies tailored on each patient was presented and evaluated on real precision dosing problems regarding both already approved and under development drugs. In this dissertation, the literature MIRL framework [39,71–74,78] was also adapted to derive a set of general clinically acceptable adaptive dosing rules (i.e., a single RL-based controller for all the patient population) on some of the proposed precision dosing scenarios to better understand its potentialities and challenges.

The novel MIRL framework implemented in this thesis relies on QL as RL algorithm and introduces a personal QL-agent (QLind) trained on each specific patient. Therefore, each patient is treated following an individually tailored RL-based adaptive dosing protocol. The central feature of this hybrid RL/PK-PD approach is the use of patient digital twin to provide the experience necessary to train the personal QL-agent. This virtual replica of the patient is represented by a PK-PD model with an individual set of parameters and covariates describing the pharmacological response for that specific individual. Such methodology allows a deeper personalization level than the literature MIRL workflow which relies on training a single RL-agent (here named QLpop-agent given the use of QL as main RL algorithm) on an entire population of patients thus learning a general adaptive dosing protocol.

Although QL is not the most recent RL algorithm (see Appendix A for a detailed description of more advanced RL techniques used for MIRL), it led to satisfactory results in all the considered case studies, thus making unnecessary to use more advanced RL methods.

Specifically, implemented QL-based patient-centric MIRL workflow was evaluated on three real precision dosing problems of increasing complexity (chapters 3-5). This assessment was characterised by some simplifications (i.e., neglection of RUV and complete knowledge of individual PK-PD model parameters) necessary to have a first clear understanding of the potentialities of the novel technique bridging RL and PK-PD modelling.

Following these assumptions, QLind-agents were first tested on the erdafitinib precision dosing problem in metastatic urothelial cancer patients (Chapter 3). This case-study was directly derived from clinical oncology where erdafitinib is administered through an adaptive dosing protocol based on the monitoring of the $[PO_4]_{serum}$, acting as efficacy/safety biomarker. The narrow therapeutic window and significant IIV of pharmacological response make erdafitinib therapy suitable to test the personalized MIRL approach. From an *in silico* evaluation on a population of 141 virtual patients, it emerged that QLind-agents outperformed both erdafitinib FDA-approved clinical protocol and the general adaptive dosing rules obtained by

QLpop-agent. Key step to obtain such results was the formalization of the clinical setup within the RL framework, including patient state representation, action selection coherent to clinical treatment management, and reward functions describing the therapeutic goal. These aspects allowed QLind agents to derive patient-specific adaptive dosing protocols maximizing treatment efficacy and avoiding severe toxicities.

Then the MIRL framework was, for the first time, evaluated on a multi-objective precision dosing scenario in which different efficacy/toxicity biomarkers have to be simultaneously optimized during pharmacotherapy (Chapter 4). In this exploration MIRL paradigm was applied with the dual goal of deriving a clinically acceptable general adaptive dosing protocol for a population of patients and individually tailored strategies. In particular, patient specific protocols were used to optimize both clinical pharmacotherapy and the drug development process by maximizing the outcomes of a clinical study. To this end, the multi-objective precision dosing problem of givinostat, a compound under clinical development for the treatment of PV, was considered as case study. Givinostat treatment inhibits the uncontrolled myeloproliferation of bone marrow cells in PV patients, and the monitoring of PLT, WBC and HCT is used to assess both therapy efficacy and toxicity as well as to guide dose adjustments. The high IIV characterizing this clinical setup makes the joint optimization of the three haematological parameters very challenging. This investigation highlighted that QLpop-agent learnt a general dosing strategy that reached similar performances to the clinical protocol approved for givinostat phase III trial. Such results confirmed the optimality of the clinical rules and highlighted the challenges of managing multiple endpoints and high IIV in a non-personalized RL-based approach. Differently, the QLind-agents successfully managed the complexity of the multiple biomarker optimization and high IIV, thus outperforming both the clinical protocol and the QLpop-agent. Thus, this analysis highlighted that, in contrast to other studies [73,79,80], in complex scenarios with multiple outcomes to simultaneously optimize, the gap between QLpop-based rules and QLind ones increases. Interestingly, this case study also remarked the flexibility of the RL framework. Indeed, by performing a simple change in the reward function, QLind-agents were able to successfully optimize the endpoints of givinostat phase III trial through patient-tailored adaptive dosing strategies.

Further, the personalized MIRL-based adaptive dosing strategy was also evaluated in the context of concomitant drug administrations to optimize first only short-term outcomes and then, short- and long-term outcomes together (Chapter 5). In particular, the complex precision dosing challenge provided by the AX-AH co-administration in advanced RCC was used to test the individual oriented MIRL approach. AX dosing is adapted based on BP monitoring, the primary efficacy/toxicity biomarker, with hypertension being a common AX-induced side effect. Treatment goal is to maximize AX dose for tumor shrinkage (SLD) and simultaneously to maintain BP within a safe range. This precision dosing problem is particularly challenging due to the significant IIV in AX response. A modeling framework describing (i) the

direct effect of AX on dBP and sVEGEFR, (ii) SLD inhibition mediated by the AX-induced reduction of sVEGEFR levels and (iii) SLD effect on patient survival probability was integrated with the MIRL framework. The sequential decision-making process characterizing the co-administration was formalized as a MDP and two different reward functions were employed. In particular, the ST-Reward was used to drive dose adjustment by focusing on short-term outcomes like BP control, increase AX exposure and limit AH usage. Conversely, the S&LT-Reward considered also long-term outcomes such as patient survival probability in the estimation of the individualized optimal dosing policy. The obtained results demonstrated that the MIRL framework was able to achieve the treatment goals also in this co-administration setting, thus confirming the powerfulness of the methodology. Although both reward functions achieved similar overall survival rates, the S&LT-based QLind agents reduced the exposure to AX as it was pushed only when a significant improvement of patient survival probability was achieved.

The results on these three challenging precision dosing problems demonstrate that when the patient PK-PD model is accurate and its parameters are fully known before treatment begins, the individual-oriented MIRL approach can effectively optimize various targeted treatment outcomes (potentially also long-term ones) for both monotherapies and concomitant administrations. However, since the underlying assumptions can prevent the application of this framework within clinical practice, two extensions of the novel individual-oriented MIRL framework were developed to overcome them (Chapter 6).

The first extension addresses the assumption that patient digital twins are fully characterized from the start of treatment, with known individual PK-PD model parameters. This hypothesis, often impractical in real-world scenarios, is mitigated by introducing a Bayesian MIRL framework. This workflow relies on Bayesian MAP estimation of patient PK-PD parameters at each monitoring step to characterize the digital twin during the treatment. Then, the current updated PK-PD model describing patient response to treatment is embedded within RL to derive a personalized set of adaptive dosing rules. This original approach was evaluated on a simulation study which used a simplified version of givinostat precision dosing problem as case study. Obtained results confirmed the potentialities of the Bayesian MIRL approach as it reached performances similar to the ideal scenario in which patient PK-PD parameters are known before treatment begin.

The second extension focuses on addressing RUV due to model misspecifications, which can affect the robustness of RL-based dosing strategies. A new QL variant, EQL algorithm, was developed to account for stochasticity in reward functions and state transitions caused by model misspecifications. During the training stage, this algorithm adopts PK-PD Monte Carlo simulations to estimate the probabilities of all potential patient states and rewards, allowing for precise updates to the estimated dosing policy. The EQL algorithm was applied to the precision dosing problem of vancomycin continuous infusion in ICU patients, a scenario characterized by

significant PK variability and the necessity for tight therapeutic monitoring. EQL was tested under varying levels of stochasticity and compared against two other agents: QLc, which considers RUV in a standard QL framework, and QLdet, which ignores RUV. The EQL-agent consistently outperformed both benchmarks, particularly under higher levels of stochasticity, thereby demonstrating its robustness in the presence of model misspecifications. Further analysis revealed that only significantly high errors in estimating RUV could decrease EQL-agent performance.

Therefore, each of the methodologies presented in Chapter 6 successfully addressed a specific limitation of the individual oriented MIRL paradigm. Although this promising result can drive MIRL closer to real-world clinical applications, some further investigations combining EQL and Bayesian MIRL are still required. In addition to these methodological aspects, further issues regarding MIRL validation should be accounted in future before clinical implementation of this methodology. Indeed, in this dissertation, similarly to literature available works [30,39,73,79,80,83–86,88,89,89,91,92], RL-based dosing strategies were evaluated through simulations. Therefore, a rigorous clinical evaluation based on randomized clinical trials will be essential to confirm that the benefits observed *in silico* can be replicated in real-world clinical settings. However, defining a clinical trial setup that meets current regulatory standards to evaluate MIRL approaches for precision dosing remains one of the most important and challenging open questions. Right now, only in one scenario, RL-based dose suggestions were applied to a small pool of healthy volunteers.

Overall, the results presented throughout this thesis highlight the potentialities of MIRL approaches to support a wide range of precision dosing tasks based on adaptive dosing protocols. Novel methods to overcome the current limitations of RL/PK-PD techniques were proposed and successful results were achieved, thus confirming the appropriateness of the developed methodologies. Although additional refinements and clinical investigations are required, the promising findings presented here provide a strong foundation for future advancements in precision dosing.

Appendix A

Overview of Reinforcement Learning Algorithms

This appendix extends the methodological dissertation on RL provided in Chapter 2. In particular, since in this thesis only QL algorithm was adopted, here the focus will be on providing an overview of the other most popular RL methods.

Generally, RL algorithms can be subdivided into two macro-categories depending on whether they require as input the Markovian state transition probability matrix (i.e., containing $p(s'|s,a)$, the probability that the system evolves to state s' at time $t+1$, given the selection of action a in state s at time t) or not. The first, is the case of the early introduced RL techniques that leveraged dynamic programming to estimate the optimal policy [76,93,94]. However, especially in the healthcare domain, the transition probability function is rarely known. Consequently, different RL algorithms were developed to solve the MDP without relying on the direct knowledge of $p(s'|s,a)$, but leveraging agent-system interplay and collecting sequences of transitions $< S_t, A_t, S_{t+1}, R_{t+1} >$. These methods, that will be covered in the following sections, are generally categorized in *value-based methods*, which rely on the estimation of state-value or action-value function, and *policy-based methods*, which directly estimate the probability distribution $\pi(s) = p(a|s)$ without learning a value-function [76,143].

A.1. Monte-Carlo Tree Search

The key of Monte Carlo Tree Search (MCTS) is to represent the entire MDP as a tree (i.e., directed graph, Figure A.1) in which the nodes are system states, S_t, and edges represent all possible actions that can be taken in S_t [76]. Therefore, the child nodes of S_t are all possible next states, S_{t+1}, that can be reached for each action (Figure A.1). Given an episode with T

decisional time steps, the states-actions path in the tree, P=$\{s_t, a_t, s_{t+1}, a_{t+1}, \ldots, a_{T-1}, s_T\}$, is called playout and it is associated with a total reward $G = \sum_t^T R_t$. MCTS combines Monte Carlo Simulation and Tree Search heuristics to find the optimal policy, $\pi^*(s) = a^*$, associating to each state the best action maximizing the final reward (i.e., the optimal path of the three) avoiding a brute force search (i.e., simulating all possible scenarios) [76].

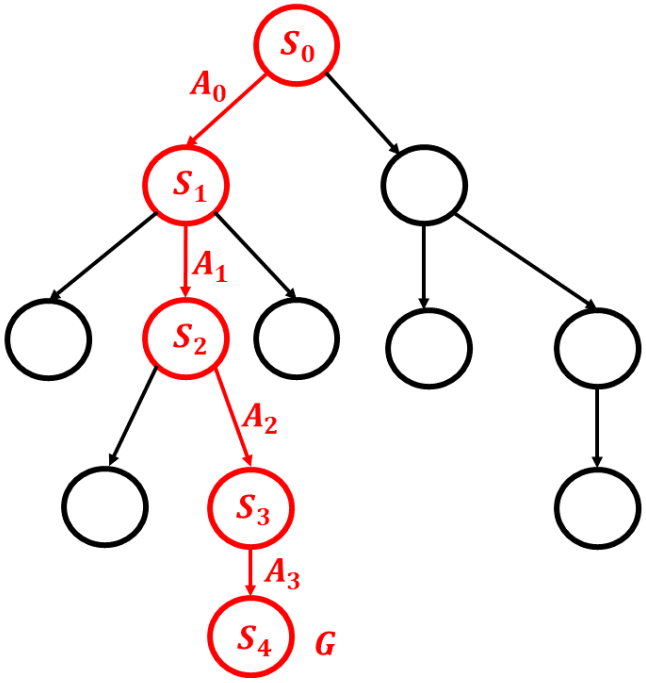


Figure A. 1: Representation of a MDP with a directed graph (tree). Each node represents the current system state, edges are the actions taken in a given system state. Each path in the tree has a score, G, representing the sum of the reward collected at each time step.

Similarly to other RL algorithms, MCTS aims to estimate the action-value function, $Q(s,a)$, for each s, a couple. Therefore, at the end of the training, for each s, the action taken by algorithm will be $a = arg\ max_a Q(s,a)$ [76].

To this end, during the algorithm training, it is necessary to store the following information for each node to estimate $Q(s,a)$ [144,145]:

- $N(s)$: the number of times in which s has been visited.
- $N(s,a)$: the number of times a was performed in s.
- $Q(s,a)$: the action-value function linked to s, a couple.

Moreover, the following policies, π^{sel} and π^{sym}, are leveraged in the training stage (Eq. A.1-A.2):

$$\pi^{sel}(s) = \ arg\ max_a \left\{ Q(s,a) + c \cdot \sqrt{\frac{2 \cdot \log\big(N(s)\big)}{N(s,a)}} \right\}$$

(A. 1)

$$\pi^{sym}(s) \ = \ a, with\ P(a|s) \ = \frac{1}{|A(s)|}.$$

(A. 2)

In particular, π^{sel} represents the upper confidence bound applied to trees (UCT) [144,145] and encapsulates the trade-off between exploration and exploitation during training. Indeed, if a given s, a couple has been explored a few times, the term under the squared root becomes predominant. The balance between exploration and exploitation is given by c representing a hyperparameter of the algorithm. Conversely, π^{sym} represents a random choice of the action since each action has a probability equal to the size of the action set in the state s, $|A(s)|$.

A comprehensive description of the MCTS algorithm is provided in the Algorithm A.1. For the sake of clarity, the main algorithm was subdivided into sub processes as described also in [146]. In particular, the algorithm estimates $Q(s,a)$ by iteratively building the search tree. At each iteration a playout is performed (i.e., simulation procedure) following π^{sel} whether S_t belongs to B, otherwise, π^{sym} is adopted and S_t is added to the tree. After the completion of the playout, the information of each node S_t are updated (i.e., backpropagation procedure). In this step, $N(S_t)$ and $N(S_t, A_t)$ are increased by 1, while $Q(S_t, A_t)$ is updated with Eq. A.3

$$Q(S_t, A_t) \ = \ Q(S_t, A_t) \ + \frac{G - Q(S_t, A_t)}{N(S_t, A_t)},$$

(A. 3)

with G being the sum of the rewards collected in the playout.

Algorithm A. 1: Pseudocode of the MCTS algorithm..

MCTS
Given set of *N* states, set of *M* actions A, hyperparameter *c*, max number of iteration *I*, T terminal time step **init** empty tree *B* **loop** for each episode (I times): $\{S_0, A_0, S_1, A_1 \ldots, A_{T-1}, S_T\}, G = \text{SIMULATION}(S_0)$ BACKPROPAGATION$(\{S_0, A_0, S_1, A_1 \ldots, A_{T-1}, S_T\}, G)$
procedure SIMULATION(S_0)
Set current state to S_0 **init** G = 0, **init** empty memory of the visited states-actions {∅}

loop until terminal state S_T is reached
    if $S_t \in B$:
        $A_t \leftarrow \pi^{sel}(S_t)$
    else
        ADDNODE(S_t)
        $A_t \leftarrow \pi^{sym}(S_t)$
    Observe next state S_{t+1} and reward R_{t+1}
    Update current system state with S_{t+1}
    $G = G + R_{t+1}$
    update the memory of the visited states-actions $\{S_0, A_0, ..., S_t, A_t\}$
return $G, \{S_0, A_0, S_1, A_1 \ldots, A_{T-1}, S_T\}$

procedure BACKPROPAGATION($\{S_0, A_0, S_1, A_1 \ldots, A_{T-1}, S_T\}, G$)

loop for $t = 0$ to $T - 1$:
    $N(S_t) \mathrel{+}= 1$
    $N(S_t, A_t) \mathrel{+}= 1$
    $Q(S_t, A_t) = Q(S_t, A_t) + \frac{G - Q(S_t, A_t)}{N(S_t, A_t)}$

procedure ADDNODE(S_t)

$N(S_t) = 0$
loop for all a in $A(S_t)$:
    $N(S_t, a) = 0$
    $Q(S_t, a) = \infty$
append S_t to B

A.2. Deep Q-Learning

Deep QL (DQL) is the natural extension of QL to a continuous domain. Indeed, DQL leverages Neural Networks (NNs) depending by a parameter set θ, i.e., $Q_{net}(\theta)$, to approximate the Q-function for each state-action pair [147]. In particular, Q_{net} is supposed to receive a vector $S \in \mathbb{R}^L$ describing the current system state as input and to return a vector $W \in \mathbb{R}^M$ containing the values of the Q-function for each of the available M-actions [96,147]. Since system state is a vector, DQL allows a deeper characterization of system condition that is not constrained to a finite set. Training a DQL agent corresponds to find the optimal set of parameters, θ^*, that best approximates the Q-function. To this end, similarly to QL, an iterative strategy in which a T-step episode is repeated I times, is performed. At each time step t, by following the ε-greedy strategy, a transition $(S_t, A_t, S_{t+1}, R_{t+1})$ is stored within a replay-buffer of size D [96]. Then, a minibatch of $D' < D$ transitions are sampled from the replay-buffer and used to update θ by performing a gradient descent step to minimize the loss function in Eq. A.4.

$$L(\theta) = \sum_i^{D'} ((R_{t+1}^{(i)} + \gamma \cdot max_A Q_{net}(S_{t+1}^{(i)}, A_{t+1}^{(i)}; \theta)) - Q_{net}(S_t^{(i)}, A_t^{(i)}; \theta))^2 \quad \text{(A. 4)}$$

DQL is more complex than QL due to the higher number of hyperparameters. Indeed, the Q_{net} architecture (i..e, number of layers, activation functions), the loss minimization algorithm with its learning rate γ and the sizes of both minibatch (D') and replay-buffer (D) should be carefully tuned [96]. Several versions of DQL were recently developed to overcome the limitations of the first implementation [96,148,149]. For example, the learning instability of $Q_{net}(\theta)$ is circumvented through the introduction of another NN, $Q_{net}{}'(\theta')$, to estimate the maximum expected return of $(S_{t+1}, max_{A_t} Q_{net}{}'(S_{t+1}, A_t; \theta'))$ [96]. $Q_{net}{}'(\theta')$ is identical to $Q_{net}(\theta)$ at the beginning of training but its parameters are updated with lag behind the regular $Q_{net}(\theta)$. Indeed, every H iterations, Q_{net}' is updated by simply performing $\theta' \leftarrow \theta$. Alternatively, Dueling-DQL (D2QL) or Dueling-Double-DQL (D3QN) were proposed to improve training stability by adopting another NN estimating the value function ($\widehat{V_{net}}$) [150,151]. The pseudocode of DQL algorithm is reported in Algorithm A.2.

Algorithm A. 2: Pseudocode of DQL algorithm.

Given: set of *M* actions, *D*-dimensional replay memory *B*, batch size *D'*, discount factor γ, a probability ϵ, a maximum number of training iterations I, update frequency *H,* probability ϵ
init two identical neural networks, Q_{net} and Q'_{net} with the same weights: $\theta'_{net} = \theta_{net}$
loop for each episode (I times):
    Set current system state to S_0
    loop for each decisional time step *t*:
        p ← uniform random number ∈ [0,1]
        if p<ϵ
            Select action A_t randomly
        else
            $A_t \leftarrow arg\ max_a\ Q_{net}(S_t, a; \theta_{net})$
        Perform A_t on the system
        Observe next state S_{t+1} and reward R_{t+1}
        Set current system state to S_{t+1}
        Store the transition tuple $(S_t, A_t, S_{t+1}, R_{t+1})$ in B
        if size of B ≥ D':
            Randomly selection of a minibatch of D ' transitions from B
            loop for each transition *i* in the of the minibatch:
                if S_{t+1} is the terminal state:
                    set $y_i = R^{(i)}{}_{t+1}$
                else:
                    set $y_i = R^{(i)}{}_{t+1} + \gamma \cdot max_a Q'_{net}(S_{t+1}, a; \theta'_{net})$
            Update the θ_{net} performing a gradient descent step on $\sum_i^K (y_i - Q_{net}(S^{(i)}{}_t, A_t{}^{(i)}; \theta_{net}))^2$
            if the number of update is multiple of *H*:
                set $\theta'_{net} = \theta_{net}$

A.3. Actor critic DQL

Actor-critic (A2C) algorithm is placed at the intersection of value-based and policy-based RL approach [96,152]. A2C merges the advantages of these two approaches by using two NNs, the actor, $\pi_{net}(\theta)$, and the critic, $V_{net}(\phi)$, characterized by their own set of learnable parameters, θ and ϕ. The actor learns $\pi(s)$ by interacting with the critic, which simultaneously estimates the appropriateness of being the system in the state s (i.e., $V^{\pi}(s)$). Analogously to QL and DQL, the training of A2C relies on repeating a T-step episode for I times. At each time step t, actions are selected by randomly sampling from actor policy, $\pi_{net}(a|s,\theta)$ and transitions $(S_t, A_t, S_{t+1}, R_{t+1}, \pi_{net}(A_t|S_t,\theta), V_{net}(S_t,\phi))$ are stored, including the probability of performing A_t given the actor's current policy and critic's evaluation of S_t. Once the episode is completed, it is possible to compute, for each S_t, the sum of the future discounted rewards by applying Eq. A.5. and then adding this information at each transition:

$$(S_t, A_t, S_{t+1}, R_{t+1}, \pi_{net}(A_t|S_t,\theta), V_{net}(S_t,\phi), G_t), \text{ with}$$

$$G_t = \sum_{j=t+1}^{T} \gamma^j \cdot r_j .$$

(A. 5)

Finally, the weights θ and ϕ are updated by performing a gradient descent step for minimizing the loss functions in Eq. A.6 and A.7.

$$Loss_{Actor} = -\sum_{t}^{T} (G_t - V_{net}(S_t,\phi)) \cdot log(\pi_{net}(A_t|S_t,\theta))$$

(A. 6)

$$Loss_{Critic} = \sum_{t}^{T} (G_t - V_{net}(S_t,\phi))^2 .$$

(A. 7)

The strategy of updating network weights after completing an episode is called Monte Carlo updating [96]. In particular, the goal of minimizing Eq. A.6 is to increase the probability of choosing action with a higher advantage, which is quantified through $G_t - V_{net}(S_t,\phi)$. Instead, minimizing Eq.A.7 is essential to update and improve the critic evaluation. Generally, π_{net} and V_{net} share the same architectural backbone (i.e., layers) as both take S_t as input. Then, this common sequence of layers bifurcates into actor and critic branches. Sometimes, computing G_t for each S_t can be computationally demanding, therefore a N-steps updating strategy can be preferred to the Monte Carlo one. In this case, at each time step t, a transition $(S_t, A_t, S_{t+1}, R_{t+1}, \pi_{net}(A_t|S_t,\theta))$ is stored in a buffer replay of size D. Then, after N temporal steps, a minibatch of D' transitions are randomly

sampled from the buffer replay. For each of them, the advantage of performing a given action in a certain state is approximated by Eq. A.8:

$$Adv = (R_{t+1} + \gamma \cdot V_{net}(S_{t+1}, \phi) - V_{net}(S_t, \phi).$$

(A. 8)

Consequently, in the N-steps strategy, Eq. A.6 and Eq. A.7 become:

$$Loss_{Actor} = -\sum_{i}^{D'} Adv^{(i)} \cdot log\left(\pi_{net}(A_t^{(i)} | S^{(i)}{}_t, \theta)\right)$$

(A. 9)

$$Loss_{Critic} = -\sum_{i}^{D'} (Adv^{(i)})^2 .$$

(A. 10)

A further extension of the presented version of A2C is adding a regularization term to Eq. A.6 or Eq. A.9 depending on the entropy of $\pi_{net}(\cdot | S_t, \theta)$ in order to improve action exploration during training [153]. A2C as well as Deep Deterministic Policy Gradient (DDPG) Evolution Strategy Algorithms are the most advanced RL algorithms which can be applied to a continuous domain of actions [152,154,155]. Pseudocodes for A2C algorithm with Monte Carlo and N-steps updating strategy are reported in Algorithm A.3 and A.4, respectively.

Algorithm A. 3: Pseudocode of Actor Critic RL with Monte Carlo update.

Given: discount factor γ, a maximum number of training iterations I
init actor network $\pi_{net}(\theta)$, critic network $V_{net}(\phi)$
loop for each episode (I times):
    Initialize a memory replay B of size T (number of decisional steps)
    Set current system state to S_0
    loop for each decisional time step t:
        Randomly select A_t from $\pi_{net}(a|S_t, \theta)$
        Perform A_t on the system
        Observe next state S_{t+1} and reward R_{t+1}
        Set current system state to S_{t+1}
        Store the tuple $(S_t, \pi_{net}(A_t|S_t, \theta), S_{t+1}, R_{t+1}, V_{net}(S_t, \phi))$ in B
    loop for each tuple in B:
        Compute the sum of discounted reward from t to T:
        $G_t = \sum_{j=t+1}^{T} \gamma^j \cdot r_j$
        Update the tuple in B appending G_t,
        $(S_t, \pi_{net}(A_t|S_t, \theta), S_{t+1}, R_{t+1}, V_{net}(S_t, \phi), G_t)$
    Update actor weights θ with a gradient descent step on
    $-\sum_{t}^{T} (G_t - V_{net}(S_t, \phi)) \cdot log(\pi_{net}(A_t|S_t, \theta))$
    Update critic weights ϕ with a gradient descent step on

$$\sum_{t}^{T}(G_t - V_{net}(S_t, \phi))^2$$

Algorithm A. 4: Pseudo code of Actor Critic RL with step update.

Given: discount factor γ, a maximum number of training iterations I, buffer replay B of size D, batch size D'
init actor network $\pi_{net}(\theta)$, critic network $V_{net}(\phi)$
loop for each episode (I times):
    Set current system state to S_0
    loop for each decisional time step t:
        Randomly select A_t from $\pi_{net}(a|S_t, \theta)$
        Perform A_t on the system
        Observe next state S_{t+1} and reward R_{t+1}
        Store the transition tuple $(S_t, \pi_{net}(A_t|S_t, \theta), S_{t+1}, R_{t+1})$ in B
        Set current system state to S_{t+1}
        if size of B $\geq D'$:
            Randomly selection of a minibatch of D' transitions from B
            loop for each transition i in the of the minibatch:
                Compute the advantage
                $Adv_i = R^{(i)}{}_{t+1} + \gamma \cdot V_{net}(S^{(i)}{}_{t+1}, \phi) - V^{(i)}{}_{net}(S_t)$
                Update the i-th tuple appending Adv_i
                $(S_t, \pi_{net}(A_t|S_t, \theta), S_{t+1}, R_{t+1}, Adv_i)$
            Update actor weights θ with a gradient descent step on
            $$-\sum_{i}^{D'}(Adv_i) \cdot log(\pi_{net}(A^{(i)}{}_t\,|S^{(i)}{}_t, \theta))$$
            Update critic weights ϕ with a gradient descent step on
            $$\sum_{i}^{D'}(Adv_i)^2$$

Appendix B

Supplementary Materials of Chapter 3

This Appendix collects all information supporting the application of the implemented MIRL approaches (sections 2.2.1. and 2.2.2.) on erdafitinib precision dosing problem presented in Chapter 3.

B.1. Erdafitinib PK-PD modelling

B.1.1. Population PK model of Erdafitinib

Erdafitinib is an orally administered drug which can be found in plasma both in free form and bounded to alpha1-glycoprotein (AGP). In particular, only the unbounded fraction of the compound can spread to peripheral tissues and biophase or be eliminated. The plasmatic free concentration of erdafitinib, C_{free} can be derived from the total one, C_{tot}, as [101]:

$$C_{free} = C_{tot} \cdot f_u$$

(B. 1)

where f_u is the fraction of unbounded drug that depends on the erdafitinib dissociation constant (K_d) and the observed concentration of AGP, $[AGP]$, in patient (Eq. B.2) [101]:

$$f_u = \frac{K_d}{K_d + [AGP]}$$

(B. 2)

In [101], the total concentration of erdafitinib, C_{tot}, was described by a three compartments pop-PK model with lagged absorption (Figure B.1, Eq. B.3). The model parametrization was based on apparent clearances and volumes. A covariate model was defined for CL_{free}/F (Eq. B.4) and $V_{2,tot}/F$ (Eq. B.5), while inter-individual variability (IIV) was considered on CL_{free}/F , $V_{2,tot}/F$, V_3/F and k_a. Residual unexplained variability (RUV) was modelled with an additive term after logarithmic transformation of both predicted and observed free and total concentrations (Eq. B.6). The parameter values of pop-PK model are reported in Table B.1, the distributions of study patient covariates [101] are summarized in Table B.2.

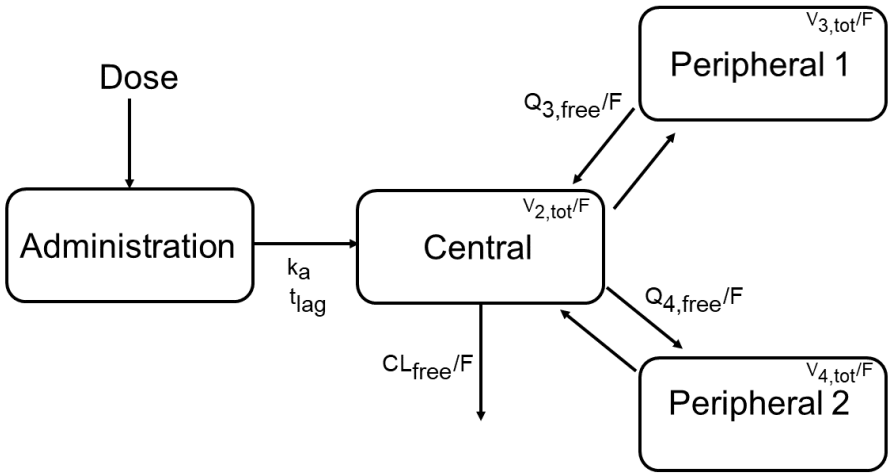


Figure B.1: Schematical representation of erdafitinib population PK model.

$$\frac{dq_{1,tot}(t)}{dt} = -k_a q_{1,tot}(t) + Dose(t)$$

$$\frac{dq_{2,tot}(t)}{dt} == k_a q_{1,tot}(t) - \frac{CL_{free}/F}{V_{2,tot}/F} f_u\, q_{2,tot}(t) +$$

$$-\frac{Q_{3,free}/F}{V_2/F} f_u\, q_{2,tot}(t) + \frac{Q_{3,free}/F}{V_3/F}\, q_3(t) +$$

$$-\frac{Q_{4,free}/F}{V_2/F} f_u\, q_{2,tot}(t) + \frac{Q_{4,free}/F}{V_4/F}\, q_4(t)$$

$$\frac{dq_3(t)}{dt} = \frac{Q_{3,free}/F}{V_2/F} f_u\, q_{2,tot}(t) - \frac{Q_{3,free}/F}{V_3/F}\, q_3(t)$$

$$\frac{dq_4(t)}{dt} = \frac{Q_{4,free}/F}{V_2/F} f_u\, q_{2,tot}(t) - \frac{Q_{4,free}/F}{V_4/F}\, q_4(t)$$

$$C_{tot}(t) = \frac{q_{2,tot}(t)}{V_{2,tot}/F}$$

$$V_2 = \frac{V_{2,tot}/F}{f_u}$$

$$q_{1,tot}(0) = q_{2,tot}(0) = q_3(0) = q_4(0) = 0$$

(B. 3)

$$\begin{cases} \left(CL_{free}/F\right)_i = \left(CL_{free}/F\right)_{pop} \exp\left(\theta_{SEX,CL} \cdot SEX_i + \right. \\ \quad +\theta_{Ren.Imp_1} Ren.Imp_{1,i} + \theta_{Ren.Imp_2} Ren.Imp_{2,i}\left.\right) \\ \quad SEX_i = 1\ if\ female, 0\ otherwise \\ Ren.Imp_{1,i}\ for\ moderate\ renal\ impairment, 0\ otherwise \\ Ren.Imp_{2,i}\ for\ mild\ renal\ impairment, 0\ otherwise \end{cases}$$

(B. 4)

$$\begin{cases} \left(V_{2,tot}/F\right)_i = \left(V_{2,tot}/F\right)_{pop} \left(\frac{[AGP]_i}{\overline{[AGP]}}\right)^{\theta_{AGP}} \exp\left(\theta_{SEX,V_2} SEX_i + \right. \\ \quad +\theta_{WT_1} WT_{1,i} + \theta_{WT_2} WT_{2,i}\left.\right) \\ \quad \overline{[AGP]}\ is\ the\ median\ of\ the\ study\ population \\ \quad SEX_i = 1\ if\ female, 0\ otherwise \\ WT_{1,i} = 1\ if\ individual\ weight < 60\ Kg, 0\ otherwise \\ WT_{2,i} = 1\ if\ individual\ weight > 80\ Kg, 0\ otherwise \end{cases}$$

(B. 5)

$$\begin{cases} C_{tot,obs} = C_{tot} \cdot \exp(\epsilon_1), with\ \epsilon_1 \sim N(0, \sigma_1^2) \\ C_{free,obs} = C_{free} \cdot \exp(\epsilon_2), with\ \epsilon_2 \sim N(0, \sigma_2^2) \end{cases}$$

(B. 6)

Table B.1: Parameter values of erdafitinib population PK model.

Parameter Name	**Description**	**Unit**	**Typical value, $\boldsymbol{\theta}$**	**Standard deviation of IIV[a], $\boldsymbol{\omega}$**
CL_{free}/F	Apparent Clearance of free drug	L/h	83.200	0.457
$Q_{3,free}/F$	Apparent inter-compartmental clearance (peripheral 1) of free drug	L/h	129.000	-
$Q_{4,free}/F$	Apparent inter-compartmental clearance (peripheral 2) of free drug	L/h	2.720	-

$V_{2,tot}/F$	Apparent volume of central compartment of total drug	L	18.900	0.230
V_3/F	Apparent volume of first peripheral compartment (only free distributes)	L	2480.000	1.050
V_4/F	Apparent volume of second peripheral compartment (only free distributes)	L	767.000	-
k_a	Absorption rate	h^{-1}	2.340	1.150
t_{lag}	Absorption lag-time	h	0.233	-
K_d	Dissociation constant to $[AGP]$ used to estimate f_u	ng/mL	32.000	-
$\theta_{SEX,CL}$	% change in CL_{free}/F for female vs male	-	-18.800	-
$\theta_{Ren.Imp_1}$	% change in CL_{free}/F for moderate renal impairment vs normal renal function	-	21.900	-
$\theta_{Ren.Imp_2}$	% change in CL_{free}/F for mild renal impairment vs normal renal function	-	21.100	-
θ_{AGP}	Power of effect of $[AGP]$ on $V_{2,tot}/F$	-	-0.473	-

θ_{SEX,V_2}	% change in $V_{2,tot}/F$ for female vs male	-	-16.400	-
θ_{WT_1}	% change in $V_{2,tot}/F$ for $< 60\ Kg$ vs $60-80\ Kg$	-	-9.820	-
θ_{WT_2}	% change in $V_{2,tot}/F$ for $> 80\ Kg$ vs $60-80\ Kg$	-	19.800	-
σ_1	Standard deviation of RUV on C_{tot}	-	0.272[a]	-
σ_2	Standard deviation of RUV on C_{free}	-	0.076[a]	-

a: standard deviation is referred to the underlying Normal distribution.

Table B.2: Statistical description of patient covariates for erdafitinib population PK model.

Continuous Covariates			
Name	**Mean**	**Standard Deviation**	**Range**
Weight [Kg]	71.60	16.60	[36.20-132.00]
[AGP] [g/L]	1.20	0.60	[0.24-3.15]
Categorical Covariates			
Name	**Values**	**Count**	**%**
Sex	Male	211	56.60
	Female	162	34.40
Renal Impairment	Normal	121	32.40
	Mild	156	41.80
	Moderate	92	25.50

B.1.2. Population PK-PD model of erdafitinib

Erdafitinib PK-PD model describes the effect of the drug on $[PO_4]_{serum}$ with respect to pre-treatment phosphate value, $[PO_4]_{BSL}$ (Eq. B.7) [100].

$$\begin{cases} [PO_4]_{serum}(t) = [PO_4]_{BSL} + M(t)C_e(t)^{\gamma} \\ \dfrac{dC_e(t)}{dt} = k_{e0}C_{tot}(t) - k_{e0}C_e(t) \end{cases} \quad (B.7)$$

A compartment effect, C_e, with a time constant k_{e0}, is introduced to account for the delay between central compartment concentration, C_{tot}, and PD effect. $M(t)$ and γ modulate the effect of C_e on $[PO_4]_{serum}$. As described in Eq. B.8 [100], , $M(t)$ is a time-varying slope which is used to describe the attenuation of drug effect during treatment:

$$\begin{cases} M(t) = m\big(1 - T(t)\big) \\ \dfrac{dT(t)}{dt} = k_{in}\big(1 - T(t)\big) \\ T(0) = 0 \end{cases}$$

(B. 8)

where T is the amount of attenuation over time, k_{in} is the rate constant describing the attenuation of drug effect with time and m is the coefficient of the slope model describing the relationship between $[PO_4]_{serum}$ and C_e at time 0. Moreover, an empirical model of $[PO_4]_{BSL}$ (Eq. B.9) was included as treatment discontinuation drives $[PO_4]_{serum}$ to a new plateau value, $[PO_4]_P$ [100]:

$$\begin{cases} [PO_4]_{BSL} = [PO_4]_0 \quad if\ T_{SLD} \le t_{lag} \\ [PO_4]_{BSL} = [PO_4]_P - ([PO_4]_P - [PO_4]_0)\exp(-k_{base}(T_{SLD} - t)) \quad otherwise \end{cases}$$

(B. 9)

where, $[PO_4]_0$ is pre-treatment baseline, T_{SLD} was the time since last dose, k_{base} was the rate of decline of phosphate baseline value with time, and t_{lag} was the time delay after which the phosphate baseline value started to decline with time. A covariate model describing the impact of sex on $[PO_4]_0$ was also developed (Eq. B.10).

$$\begin{cases} [PO_4]_{0,i} = [PO_4]_{0,pop} \cdot \exp(\theta_{SEX} \cdot SEX_i) \\ SEX_i = 1\ if\ female, 0\ otherwise \end{cases}$$

(B. 10)

IIV was introduced on k_{e0}, m and k_{in} by using a log-normal distribution, while $[PO_4]_0$ and $[PO_4]_P$ were assumed to be normally distributed across the population. Furthermore, a correlation between k_{e0} and $k_{in}$was defined in the model. The other PD parameters were fixed to their population value. Analogously to the Pop-PK model, RUV on $[PO_4]_{serum}$ was assumed log-normally distributed (Eq. B.6), with σ_3 standard deviation of the underlying Normal distribution. Table B.3 summarizes the estimated values for erdafitinib Pop-PK-PD parameters.

Table B. 3: Parameter values of erdafitinib Population PK-PD model.

Parameter Name	**Description**	**Unit**	**Typical value, $\boldsymbol{\theta}$**	**Standard deviation of IIV, $\boldsymbol{\omega}$**
$[PO_4]_0$	Phosphate baseline	mg/dL	3.080	0.463[b]
m	Slope of drug effect	$(mg/dL(ng/mL))$	0.869	0.394[a]
γ	Exponent of drug effect	-	0.860	-
k_{e0}	Effect compartment rate	h^{-1}	0.199×10^{-1}	0.909[a]
k_{in}	Rate of drug effect attenuation over time	h^{-1}	0.102×10^{-3}	1.150[a]
$[PO_4]_P$	Post-treatment interruption phosphate baseline plateau	mg/dL	2.670	0.498[b]
k_{base}	Rate of change over time for post-treatment interruption phosphate baseline plateau	h^{-1}	0.101×10^{-1}	-
t_{lag}	Post-treatment interruption lag-time before phosphate baseline value decline	h	143.000	-
θ_{SEX}	% change in $[PO_4]_0$ for female vs male	%	12.400	-

$\rho_{k_{in}-k_{e0}}$	Correlation between k_{in} and k_{e0}	-	0.577	-
σ_3	Standard deviation of RUV on $[PO_4]_{serum}$	-	0.119[a]	-

a: standard deviation refers to the underlying Normal distribution.
b: standard deviation refers to the Normal distribution.

B.1.3. Model analysis

A model analysis was performed to characterize *a priori* the response to erdafitinib treatment for each patient. In particular, it allows to compute the individual therapeutic window, i.e., the dose range that brings $[PO_4]_{serum}$ in $[5.5,7)mg/dL$ at day 140 (i.e., five months) $[100,107]$. Therefore, it was assumed a time $t \to 140\ days$ and that drug concentration in the effective compartment, C_e, was in a steady state equilibrium with the concentration in central compartment, C_{tot} (Eq. B.11). A multiplicative factor of 1000/24 was introduced in Eq. B.11 as drug concentrations are measured in $[ng/L]$ while the dose, D_{ss}, and the volume term in CL_{free}/F, were expressed in mg and L/h, respectively.

$$C_{tot} = C_e = \frac{D_{ss}}{(CL_{free}/F) \cdot 24} \cdot 1000$$

(B. 11)

The analytical expression of $T(t)$ (Eq. B.12) was obtained by solving the ordinary differential equation in Eq. B.8:

$$T(t) = 1 - \exp(-k_{in} \cdot t).$$

(B. 12)

Then, by substituting Eq. B.12 in the first equation in Eq. B.8, the formula of $M(t)$ is obtained. In particular, $k_{in}\ [h^{-1}]$ was multiplied by 24 for converting it into $[days^{-1}]$ (Eq. B.13):

$$M(t) = m \cdot \exp(-k_{in} \cdot 24 \cdot t)$$

(B. 13)

By placing Eq. B.11 and Eq. B.13 in the first equation of Eq. B.7, it is possible to derive the expression of $[PO_4]_{serum}$ as function of D_{ss} (Eq. B.14). By inverting this relationship, for $t = 140\ days$, it is possible to derive Eq. B.15 which describes the dose needed for achieving a given $[PO_4]_{serum}$ after five months of treatment.

$$[PO_4]_{serum} = [PO_4]_{BSL} +$$

$$+m \cdot \exp(-k_{in} \cdot 24 \cdot t) \cdot \left(\frac{D_{ss}}{(CL_{free}/F) \cdot 24} \cdot 1000\right)^{\gamma}$$

(B. 14)

$$D_{ss}([PO_4]_{serum}) = \sqrt[\gamma]{\frac{[PO_4]_{serum} - [PO_4]_{BSL}}{m \cdot \exp(-k_{in} \cdot 24 \cdot 140)}} \cdot \frac{(CL_{free}/F) \cdot 24}{1000}$$

(B. 15)

As will be discussed in section B.2, the therapeutic window on a five-months treatment can be computed for each patient, by applying Eq. B.15 with the individual parameters for the two bounds of the phosphate normality range (i.e., $[PO_4]_{serum} = 5.5\ mg/dL$ and $[PO_4]_{serum} = 7\ mg/dL$).

B.1.4. Mlxtran code for erdafitinib PK-PD model

```
DESCRIPTION:

[LONGITUDINAL]
input = {tlag, ka, Cl, V2, Q3, V3, Q4, V4, fu, ke0,
gamma, m, kin, PO40, PO4P, kbase, tlag_att, q10, q20,
q30, q40, T0, Ce0, TSLD}
TSLD={use=regressor}

PK:

V2tot=V2
V2free=V2tot/fu
k02 =Cl/V2tot
k32= Q3/V2free                          ;inter-compartment
clearance
k23=Q3/V3                               ;inter-compartment
clearance
k42=Q4/V2free                                ;inter-
compartment clearance
k24=Q4/V4                               ;inter-compartment
clearance
k02=Cl/V2tot                            ;elimination
clearance

depot(type=1, target=q1, Tlag=tlag)

EQUATION:
; PK model definition
;t0=0

;odeType = stiff
```

```
q1_0=q10
q2_0=q20
q3_0=q30
q4_0=q40

ddt_q1=-ka*q1
 ;administration compartment
ddt_q2=ka*q1-k02*fu*q2-k32*fu*q2+k23*q3-k42*fu*q2+k24*q4
 ;central compartment
ddt_q3=k32*fu*q2-k23*q3
 ;peripheral compartment 1
ddt_q4=k42*fu*q2-k24*q4
 ;peripheral compartment 2

Ctot=q2/V2tot
Cfree=fu*Ctot

;PD model definition
Ce_0=Ce0
T_0=T0

ddt_Ce=ke0*Ctot-ke0*Ce
ddt_T=kin*(1-T)


M=m*(1-T)

if TSLD>tlag_att
 PO4_BSL=PO4P-(PO4P-PO40)*exp(-kbase*(TSLD-tlag_att))
else
 PO4_BSL=PO40
end


PO4=PO4_BSL+M*(Ce^gamma)

;reward definition
xc=6.25

lambda1=0.5
lambda2=1.5

max_reward_range_target=1
min_reward_range_target=0.5

slope=1
reward_init=0

if(PO4>=5.5 & PO4<xc)
       intercept1=min_reward_range_target
       reward=min(1,slope*(PO4-5.5)+intercept1)
elseif (PO4>=xc & PO4<7)
 intercept2=max_reward_range_target
       reward=min(1,-slope*(PO4-6.5)+intercept2)
elseif (PO4>=7 & PO4<=9)
       reward=min_reward_range_target*exp(-lambda1*(PO4-
7))
```

```
elseif PO4>9
 reward_init=min_reward_range_target*exp(-lambda1*(9-7))
       reward=reward_init*exp(-lambda2*(PO4-9))
elseif PO4<5.5
 reward=min_reward_range_target*exp(-lambda2*(-(PO4-
5.5)))
end

OUTPUT:
output={Ctot, Cfree, PO4, Ce, T, q1, q2, q3, q4, reward}
```

B.2. Virtual patient population to evaluate the RL/PK-PD approach

B.2.1. Statistical distributions of covariates

The following statistical assumptions were made for extracting plausible covariate values given the data available in the literature (Table B.2):

- Categorical covariates were randomly sampled from both binomial (Sex) and multinomial distribution (Renal impairment) according to frequencies in Table B.2.
- Continuous covariates (i.e., weight and $[AGP]$) were extracted from a log-normal distribution with mean and standard deviation fixed to values in Table B.2. Samples outside the ranges reported in Table B.2 were fixed to boundary values.

B.2.2. Generation of the preliminary virtual population

The following steps describe the operations performed to obtain a preliminary pool of virtual patients. From this cohort, the final virtual patient population used for erdafitinib case study, was Then, by applying the filtering criteria described in next sections, the virtual population of erdafitinib patients was obtained.

- Given the distributions of PK-PD model parameters (sections B.1.1. and B.1.2.) and patient covariates (section B.2.1.) an initial pool of 100000 virtual patients was randomly extracted.
- For each patient, individual parameters were replaced into Eq.S2.15 which was evaluated with $[PO_4]_{serum} = 5.5\, mg/dL$ and $[PO_4]_{serum} = 7\, mg/dL$ to compute the therapeutic range (i.e., lower and maximum daily dose level leading $[PO_4]_{serum}$ to the target range after five months of treatment).
- Patients were subdivided in the following groups according to the computed therapeutic range:

- **Completely responsive patients** (41%): if the lower bound of the therapeutic window is $\leq 9\, mg/day$ (i.e., maximum erdafitinib dose level).
- **Not completely responsive patients** (59%): if the upper bound of the therapeutic window is $> 9\, mg/day$ (i.e., maximum erdafitinib dose level).

B.2.3. Inclusion criteria for completely responsive patients

Completely responsive patients were subdivided based on the discrete erdafitinib doses (i.e., 4, 5 6, 8, 9 [mg/day]) that fall inside the individual therapeutic ranges, as reported in Table B.4.

Table B.4: Classes of responsive patients according to the erdafitinib doses falling the individual therapeutic window

Optimal Doses [mg/day]	**% of patients in Completely responsive group**
< 4	5.130
4	4.910
5	0.000
6	0.048
8	0.000
9	12.475
4,5	6.217
5,6	4.463
6,8	1.536
8,9	27.890
4,5,6	10.790
5,6,8	5.252
6,8,9	13.545
4,5,6,8	1.548
5,6,8,9	5.349
4,5,6,8,9	0.839

Then, a five-month treatment was simulated for each patient using the personal optimal dose levels in Table B.4. The first group (i.e., Optimal dose < 4 mg/day) was not considered in this analysis as none of erdafitinib discrete dosages fall within it. These simulations confirmed a high IIV in treatment response. As illustrated in Figures B2-B4, although sharing the same optimal doses, some patients undergo hyperphosphatemia, while others maintain $[PO_4]_{serum}$ under toxicity threshold.

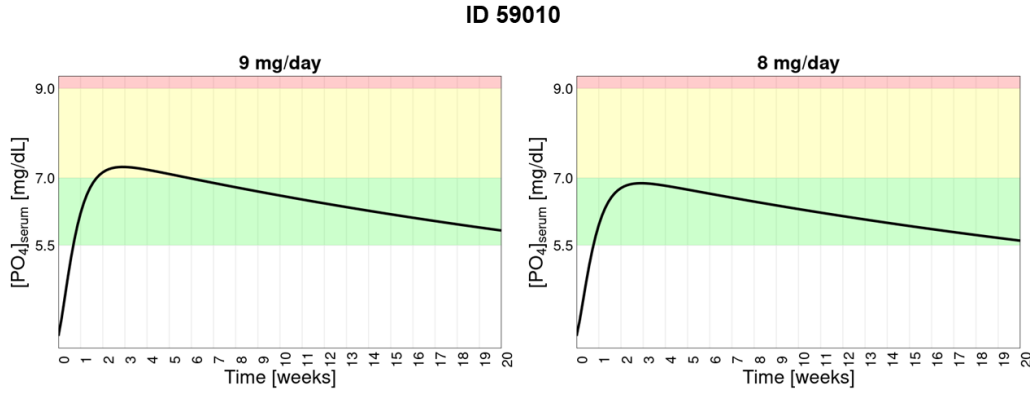


Figure B.2: Example of patient for which only 8 [mg/day] normalizes $[PO_4]_{serum}$ without leading to hyperphosphatemia.

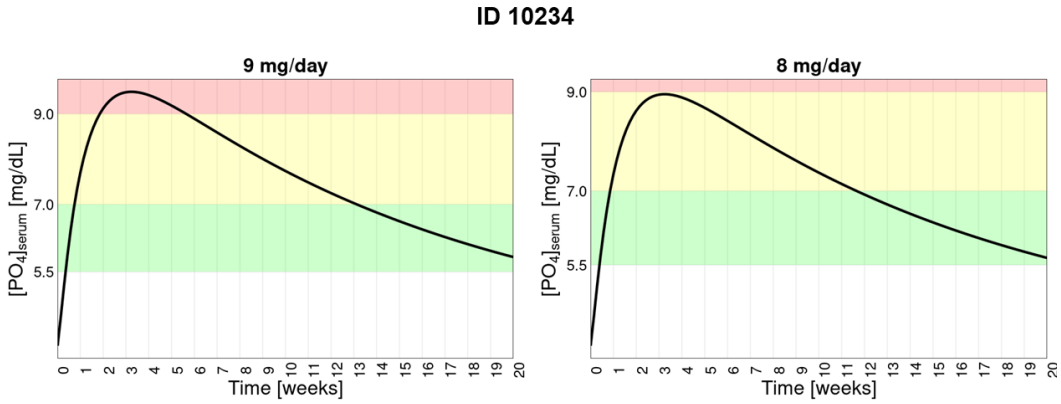


Figure B.3: Example of patient for which both 8 and 9 [mg/day] normalize $[PO_4]_{serum}$ with a hyperphosphatemia event.

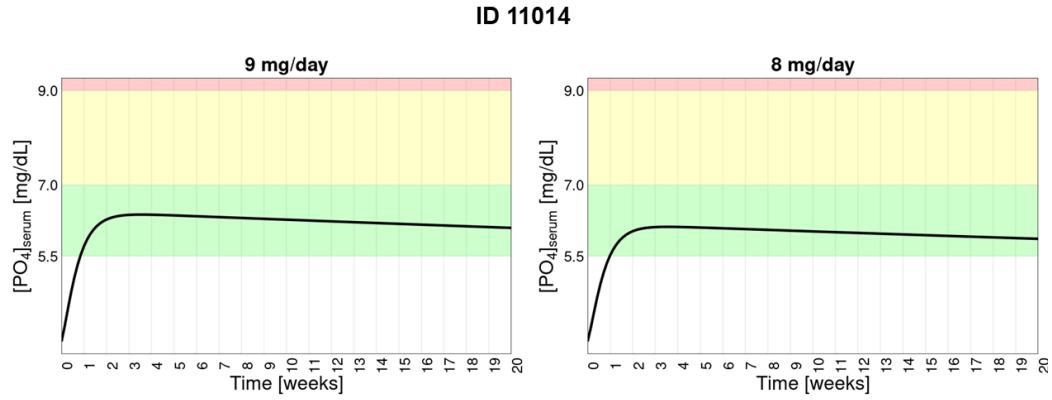


Figure B.4: Example of patient for which both 8 and 9 [mg/day] normalize $[PO_4]_{serum}$ without any hyperphosphatemia event.

Thus, the groups in Table B.4 were further split to account for this variability in the virtual population. In particular, given the ordered set of optimal doses that identifies the individual class, $D = \{d_1, \cdots, d_N\}$, the $N + 1$ subgroups S_j were defined:

- if $N > 1,\ \forall\, j = 1, \dots, N-1,\ S_j$ includes patients for which the doses in $D_j = \{d_k\}$ with $k = 1, \dots, j$, keep $[PO_4]_{serum}$ in target without leading to hyperphosphatemia. Therefore, subjects in S_j undergo hyperphosphatemia during treatment with a constant dose level in $D - D_j$.
- S_N includes the subjects having hyperphosphatemia $\forall\, d_i \in D$.
- S_{N+1} includes the subjects not having hyperphosphatemia $\forall\, d_i \in D$.

An example of this stratification is reported in Table B.5.

Table B.5: Example of the stratification strategy for patients with $[PO_4]_{serum}$ normalized at 140th day by both 8 and 9 mg/day.

Original Group	
$[PO_4]_{serum}$ in target range at 140th day with both 8 and 9 [mg/day]	
Subgroup	**Description**
S1	Patients for which only 8 [mg/day] normalizes $[PO_4]_{serum}$ without leading to hyperphosphatemia (Figure B.2).
S2	Patients for which both 8 and 9 [mg/day] normalize $[PO_4]_{serum}$ with a hyperphosphatemia event (Figure B.3).
S3	Patients for which both 8 and 9 [mg/day] normalize $[PO_4]_{serum}$ without any hyperphosphatemia event (Figure B.4).

Table B.6 summarizes the resulting subdivision of groups in Table B.4 based on the presence of hyperphosphatemia events. For each of these classes, two subjects were randomly selected, thus obtaining a pool of 92 virtual responsive patients. 4 more patients (i.e., equal to the minimum sample size in Table B.6) were randomly selected from the patients having the therapeutic window under 4 mg/day (i.e., first group in Table B.4). Therefore, the number of responsive patients in the population used for evaluating QL agents totals to 96.

Table B.6: Final stratification and sampling strategy of responsive patients.

Optimal Doses [mg/day]	Number of Subgroups	Number of randomly sampled patients
4	2	4
5	2	0[a]
6	2	4
8	2	0[a]
9	2	4
4,5	3	6
5,6	3	6
6,8	3	6
8,9	3	6
4,5,6	4	8
5,6,8	4	8
6,8,9	4	8
4,5,6,8	5	10
5,6,8,9	5	10
4,5,6,8,9	6	12

a: No patients were sampled from this group as the number of subjects belonging is 0

B.2.4. Inclusion criteria for partially responsive patients

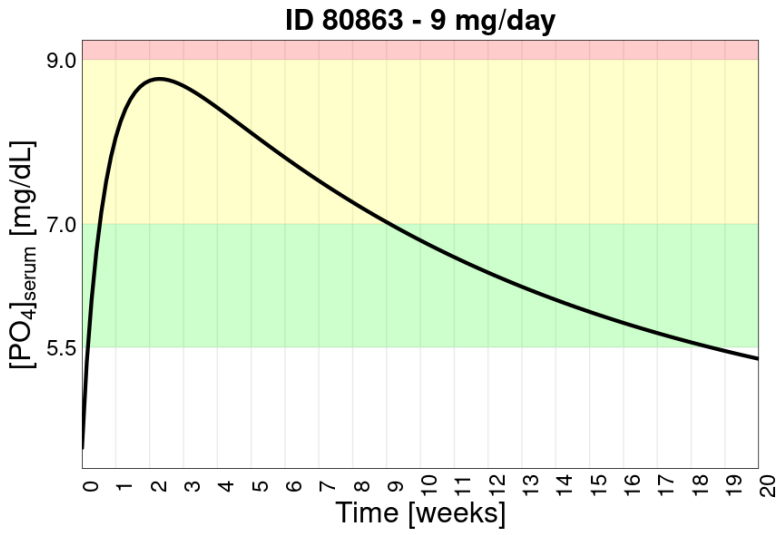


Figure B.5: Example of a partially responsive patient

By observing the response of a five-months treatment at maximum tolerated dose (i.e., 9 mg/day), the group of unresponsive patients (59% of total population) was initially stratified in:

- **Partially responsive patients** (22% of the total population): if $[PO_4]_{serum}$ can be temporarily maintain in target but, at 140th day, it is outside the range [5.5,7) mg/dL as erdafitinib efficacy decreases over time (sections B.1.2. and B.1.3.). Figure B.5 illustrates $[PO_4]_{serum}$ profile of a partially responsive patient.
- **Untreatable patients** (37% of the total population): if $[PO_4]_{serum}$ is always in the inefficacy range (i.e., $< 5.5\ mg/dL$).

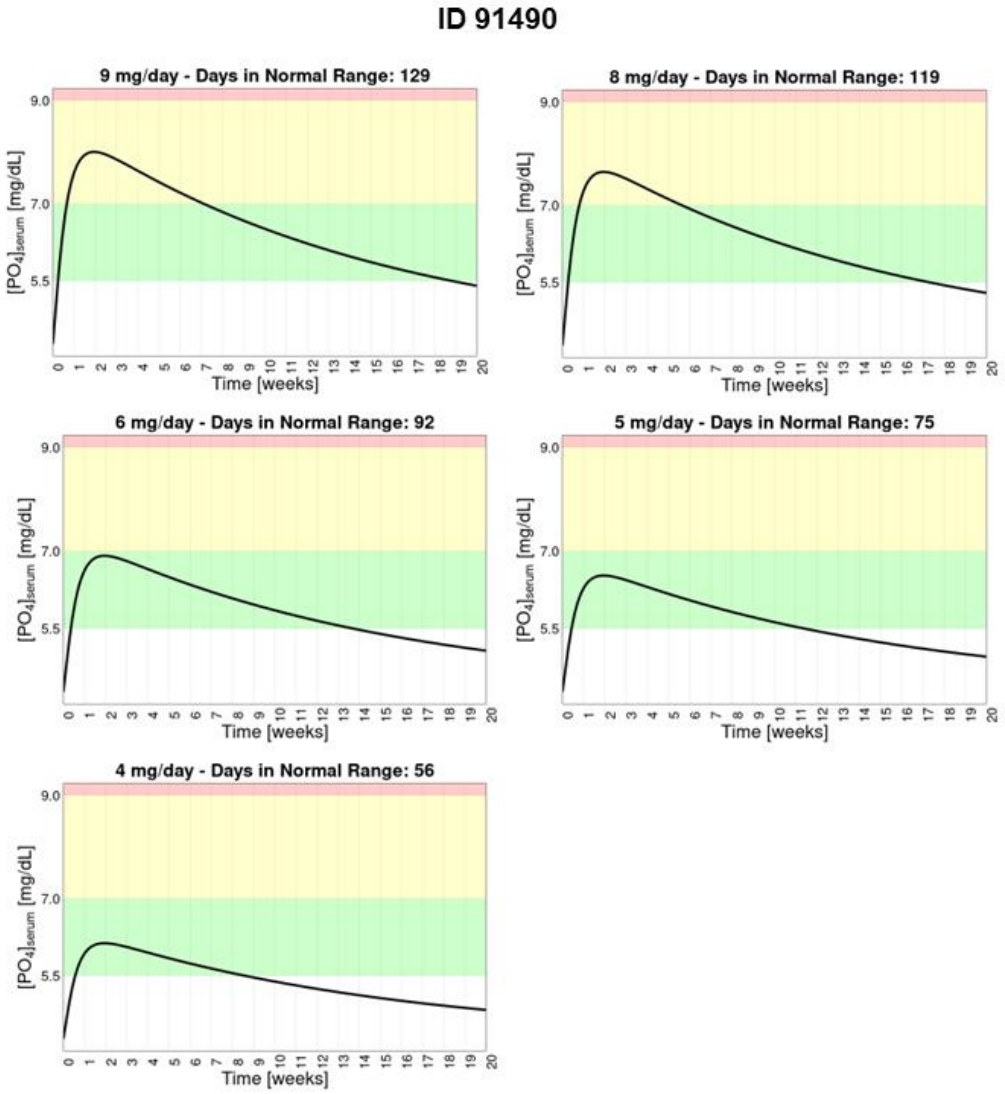


Figure B.6: Example of how partially responsive patients are stratified according to their optimal dose. In this case, the best dosage level is 6 mg/day as it keeps $[PO_4]_{serum}$ within the normality range for longer time without any hyperphosphatemia event.

Under the assumption that an ineffective drug would never be selected for the treatment of a certain disease, untreatable patients were not considered in the final population. Conversely, 45 partially responsive patients were included in the target population after the further stratification in Table B7. In such case, the optimal dose level is the amount of drug that keeps for

longer time $[PO_4]_{serum}$ in the target range without any hyperphosphatemia event. This characterization was performed by considering for each partially responsive patient, a five-months treatment at each available discrete erdafitinib dose level. As the partially responsive patients are less than the completely ones in the whole population, 45 subjects were randomly selected from the classes in Table B7.

Table B.7: Summary of the stratification within the group of partially responsive patients.

Optimal Dose level [mg/day]	**Total Count**	**Number of randomly sampled patients**
4	17909	9
5	1266	9
6	1522	9
8	382	9
9	179	9

B.3. Hyperparameters of QL algorithm

Table B.8 reports the hyperparameters used to train both the individual QL-agents and the single QL-agent to derive a general adaptive dosing protocol in the population.

Table B.8: QL hyperparameters adopted in erdafitinib case study. Such values were used both to derive patient-specific adaptive dosing protocols and to obtain a general set of dosing rules in the population.

Hyperparameter	**Value**
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.98
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{15000}\right)\right)\right)$ with i being the current iteration number

Appendix C

Supplementary Materials of Chapter 4

This Appendix contains further information supporting the application of the developed MIRL approaches (sections 2.2.1. and 2.2.2.) on givinostat precision dosing problem presented in Chapter 4.

C.1. Terms in the reward function evaluating the derivative of the haematological parameters

Figures C.1-3 provide a representation of the terms in the reward function (Eqs. 23-24) based on the derivative, y', of the haematological parameters (i.e., $Reward_{PLT,Der}$, $Reward_{WBC,Der}$, $Reward_{HCT,Der}$, Eqs. 31-33). In particular, representation of $Reward_{,Der}$ is stratified according to the efficacy (Figure C.1), toxicity (Figure C.2) and inefficacy (Figure C.3) range in which the haematological parameter falls. As in this case study a toxicity range was not defined for HCT, only functions in Figure C.1 and Figure C.3 were used for this haematological parameter.

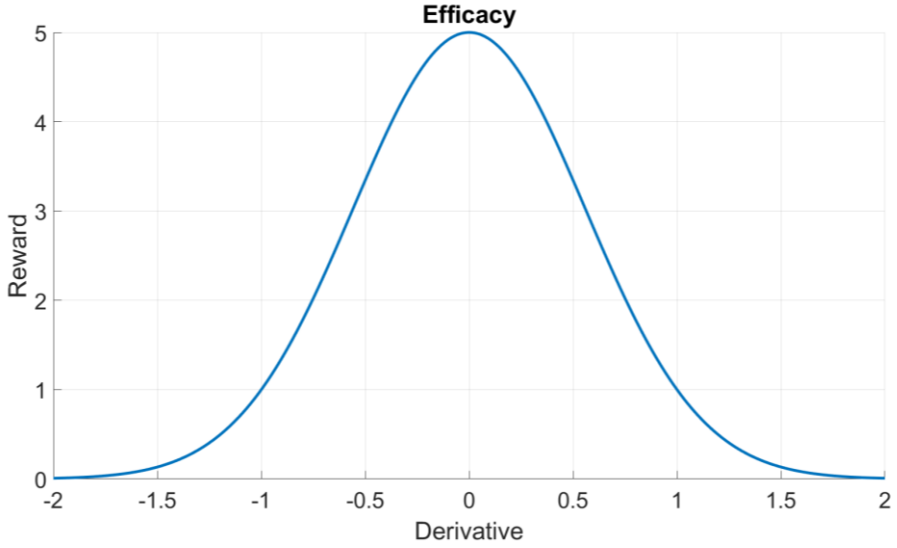


Figure C.1: Reward function adopted to evaluate the derivative of PLT, WBC and HCT when their values fall within the efficacy range (i.e., $PLT \in [150{,}400] \times 10^9/L, WBC \in [4{,}10] \times 10^9/L, HCT < 45\%$). Values are reported in the [0,5] scale accordingly to weights in Eq. 24.

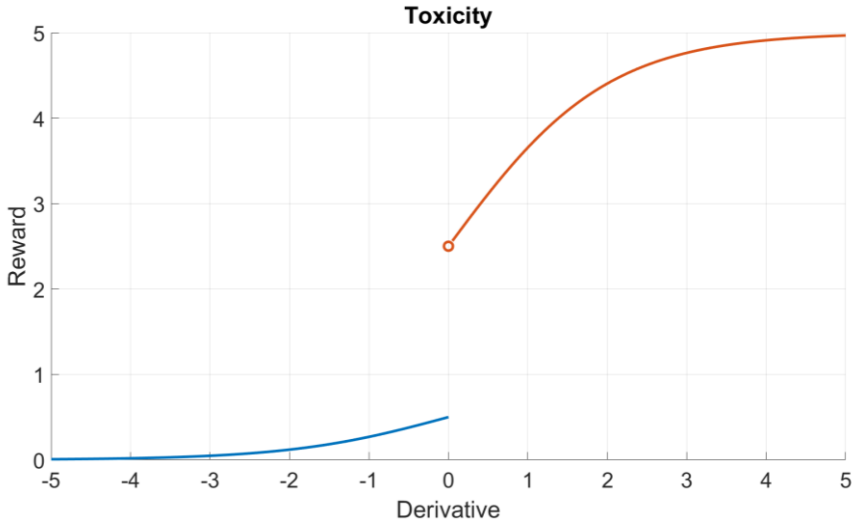


Figure C.2: Reward function adopted to evaluate the derivative of PLT and WBC when their values fall within the toxicity range (i.e., $PLT < 150 \times 10^9/L, WBC < 4 \times 10^9/L$). This function was not applied to HCT derivative as a toxicity range was not defined for this parameter. Values are reported in the [0,5] scale accordingly to weights in Eq. 24.

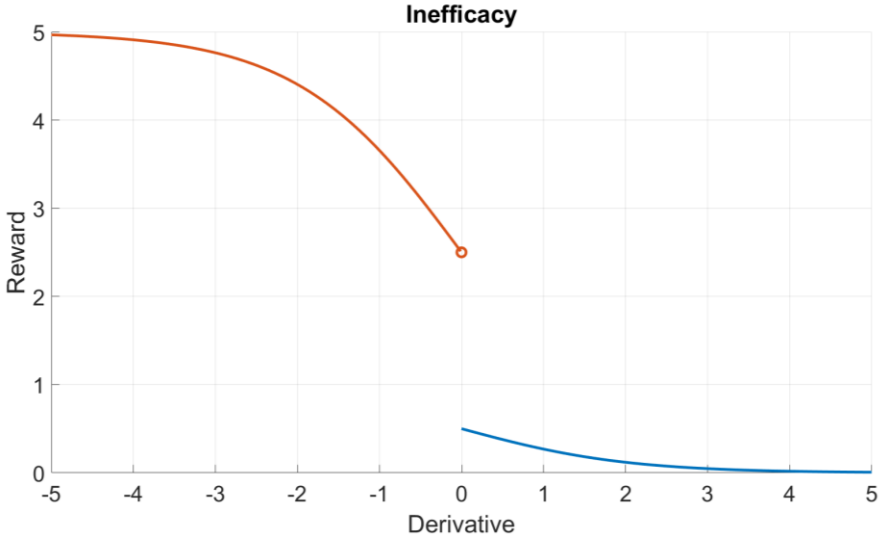


Figure C.3: Reward function adopted to evaluate the derivative of PLT, WBC and HCT when their values fall within the inefficacy range (i.e., $PLT > 400 \times 10^9/L, WBC > 10 \times 10^9/L, HCT \geq 45\%$). Values are reported in the [0,5] scale accordingly to weights in Eq. 24.

C.2. Givinostat PK-PD modelling framework

C.2.1. Population model of givinostat PK

The PK model for givinostat is a two compartmental model with a lagged first order absorption for oral administration and elimination from central compartment [45]. Table C. 1 reports the values of model parameters. A Log-Normal distribution was assumed for all the model parameters. Givinostat clearance (CL) is influenced by individual body weight scaled by median weight in the population which was equal to 77 Kg (Eq. C.1).

$$CL_i = \left(\frac{Weight_i}{77}\right)^{\theta_{Weight}} \quad (C.1)$$

Table C. 1: Parameter values of Givinostat population PK model.

Parameter	Description	Unit	Value	Variance of IIV, Ω^2
CL/F	Apparent clearance	L/h	181	0.083
V_2/F	Apparent volume of distribution of central compartment	L	171	0.490

Q/F	Apparent inter-compartment clearance	L/h	33.3	-
V_3/F	Apparent volume of distribution of peripheral compartment	L	491	0.070
K_A	Absorption rate	1/L	0.233	0.058
ALAG	Lag time	h	0.221	-
θ_{Weight}	Body Weight covariate effect	-	0.402	-

C.2.2. PK-PD modelling of givinostat effect on PLT, WBC and HCT

Givinostat myelosuppressive effect on blood cells was described by a joint Friberg model for PLT, WBC and HCT as illustrated in Figure C.4 and Eq. C.2 [45,156]. The values of model parameters are reported in Table S2.2.

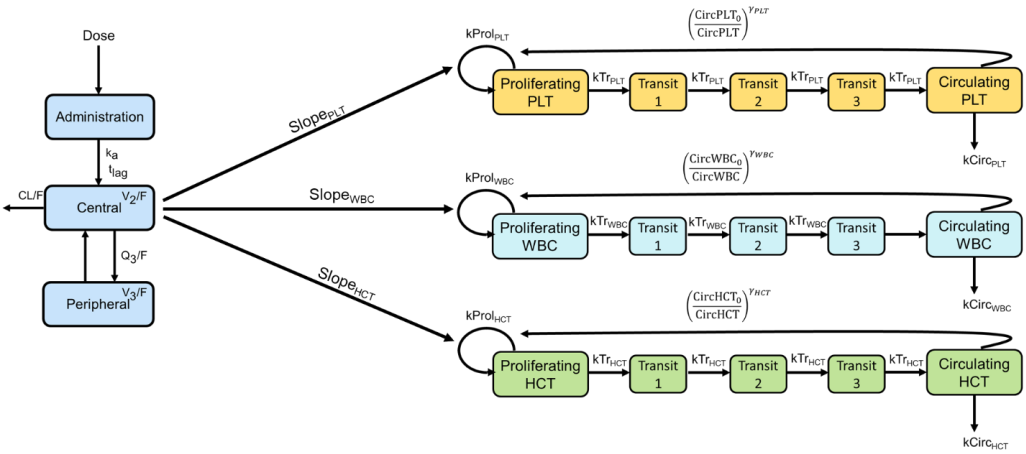


Figure C.4: Schematical representation of givinostat population PK-PD model.

$$\frac{dProl(t)}{dt} = k_{prol} \cdot Prol(t) \cdot \left(1 - E_{drug}(t)\right) \cdot \left(\frac{Circ_0}{Circ(t)}\right)^{\gamma} - k_{tr} \cdot Prol(t)$$

$$\frac{dTransit_1(t)}{dt} = k_{tr} \cdot Prol(t) - k_{tr} \cdot Transit_1(t)$$

$$\frac{dTransit_2(t)}{dt} = k_{tr} \cdot Transit_1(t) - k_{tr} \cdot Transit_2(t)$$

$$\frac{dTransit_3(t)}{dt} = k_{tr} \cdot Transit_2(t) - k_{tr} \cdot Transit_3(t)$$

$$\frac{dCirc(t)}{dt} = k_{tr} \cdot Transit_3(t) - k_{circ} \cdot Circ(t)$$

$$E_{drug}(t) = Slope \cdot conc_p(t)$$

$$Prol(0) = Transit_1(0) = Transit_2(0) = Transit_3(0) = Circ_0$$

$$k_{tr} = k_{prol} = k_{circ} = \frac{N+1}{MTT},$$

with N = number of transit compartment = 3.

(C. 2)

Table C.2: Parameter values of givinostat population PK-PD model.

Parameter	**Description**	**PLT**	**WBC**	**HCT**
Value				
MTT	Mean Transit Time (h)	297	319	610
$Circ_0$	Steady state circulating level (10^9/L for PLT and	670	15	46.1
γ	Feedback	0.142	0.097	0.21
Slope	Drug potency (L/ng)	3.74	1.94	0.313
Inter-Individual Variability (IIV), Ω				
Omega MTT	Variance (CV%)	0.085 (29.7)	0.017 (13.0)	0.095 (31.5)
Omega $Circ_0$	Variance (CV%)	0.082 (29.2)	0.118 (35.3)	0.002 (4.78)
Omega γ	Variance (CV%)	0.167 (42.6)	0.079 (28.6)	0 fixed
Omega Slope	Variance (CV%)	0.199 (47.0)	0.142 (39.0)	1.27 (160)

Correlation Between IIV							
Omega (2,1)	Covariance (Correlation) MTT_{PLT}/MTT_{WBC}	0.038 (1.000)		---	---	---	---
Omega (8,7)	Covariance (Correlation) $\gamma_{PLT}/\gamma_{WBC}$	0.1 (0.871)	49.0	---	---	---	---
Omega (11,10)	Covariance (Correlation) $Slope_{PLT}/Slope_{WBC}$	0.106 (0.628)	43.0	---	---	---	---
Omega (12,10)	Covariance (Correlation) $Slope_{PLT}/Slope_{HCT}$	0.238 (0.473)	49.7	---	---	---	---
Omega (12,11)	Covariance (Correlation) $Slope_{WBC}/Slope_{HCT}$	0.225 (0.530)	43.0	---	---	---	---

C.2.3. Steady-state analysis of givinostat PK-PD model

A steady-state analysis of the joint PK-PD Friberg model was performed to characterize *a priori* the response to givinostat treatment for each patient. Assuming a constant concentration for givinostat, i.e., $conc_p(t) = \bar{c}$, the equilibrium points of the system $\left(\overline{Prol}, \overline{Transit_1}, \overline{Transit_2}, \overline{Transit_3}, \overline{Circ}\right)$ can be derived zeroing the differential equations in Eq. C2. It follows that they have to satisfy the relationships:

$$\overline{Prol} = \overline{Transit_1} = \overline{Transit_2} = \overline{Transit_3} = \overline{Circ}$$

(C. 3)

$$\left(\frac{Circ_0}{\overline{Circ}}\right)^{\gamma} = 1 - \bar{c} \cdot Slope.$$

(C. 4)

These relationships hold for PLT, WBC and HCT. Considering the daily exposure of givinostat, i.e., $AUC_{0-24h} = \bar{c} \cdot 24h = \overline{Dose}/CL$, we can express $\bar{c}$ as $\bar{c} = \overline{Dose}/(CL \cdot 24h)$, and, replacing it in Eq. C.4, we obtain:

$$\overline{Dose} = CL \cdot 24h \cdot \frac{\left[1 - \left(\frac{Circ_0}{\overline{Circ}}\right)^{\gamma}\right]}{Slope}.$$

(C. 5)

Therefore, given individual PK-PD parameters and the target range of each biomarker, Eq. C.5 can be applied to compute the theorical dosing window leading to the complete haematological response. As will be discussed in *Supplementary Materials S3*, this result will be adopted to define the virtual population on which the methodology will be applied.

C.3. Generating a virtual population of Polycythemia Vera patients

The aim of this section is to provide a comprehensive description of the steps followed to generate the virtual populations of PV patients used as training and test sets for the QL-agents. As illustrated in Figure 25, the QLpop agent was trained on a pool of 98 virtual patients and then evaluated on 10 test sets of different PV patients sharing the same givinostat PK-PD response dynamics. Therefore, each test set and the training virtual population were generated using the procedures illustrated in this section. Differently, as detailed in section 4.1.2.6, the QLind-agents were applied to tailor the treatment of the individuals in QLpop-training set.

To generate a plausible virtual population of PV patients, it was necessary to assume a probability distribution for the only covariate included in the givinostat PK-PD i.e., body weight (WT) affecting the givinostat CL (Eq. C1). In particular, it was hypothesized that $WT \sim (77,16^2)$ according to distribution of body weight in PV patients on which the model was originally built [45]. Then, the following steps were implemented:

- Given the distributions of PK-PD model parameters (section C.2) and patients body weight WT an initial pool of 500000 virtual patients (i.e., 500000 sets of individual PK-PD model parameters and WT) was randomly extracted.
- For each virtual patient a theoretical dose window was defined for PLT, WBC and HCT with Eq. C2. In particular, lower and upper bounds of therapeutic ranges were obtained by replacing $\overline{Circ}$ with the efficacy range limits (i.e., $PLT \in [150,400] \times 10^9/L$, $WBC \in [4,10] \times 10^9/L$, $HCT < 45\%$).
- The three dose ranges obtained for PLT, WBC and HCT, were used to classify PV patients in:
    - **Theoretically responders** (57.10%): if the intersection of the dose windows of PLT, WBC and HCT is not-empty and includes levels lower than the maximum tolerated dose of 200 mg/day.

- **Theoretically not responders** (42.90%): if the intersection of the PLT, WBC and HCT dose windows:
  - is empty;
  - is not empty but it includes only dose higher than 200 mg/day;

- Under the assumption that an ineffective drug would never be selected for the treatment of a certain disease, only theoretically responders were considered for the analysis.
- Theoretically responders were subdivided into 14 groups (Table S3.1) based on which available givinostat doses are included in the intersection range previously defined.

Table C.3: Groups defined according to givinostat theoretical optimal dose.

Group	**Description**
<50	The upper bound of the intersection range is below 50 mg/day.
50	50 mg/day is the only dose falling within the intersection range.
75	75 mg/day is the only dose falling within the intersection range.
100	100 mg/day is the only dose falling within the intersection range.
125	125 mg/day is the only dose falling within the intersection range.
150	150 mg/day is the only dose falling within the intersection range.
175	175 mg/day is the only dose falling within the intersection range.
200	200 mg/day is the only dose falling within the intersection range.
50,75	50 and 75 mg/day are the only doses falling within the intersection range.
75,100	75 and 100 mg/day are the only doses falling within the intersection range.
100,125	100 and 125 mg/day are the only doses falling within the intersection range.
125,150	125 and 150 mg/day are the only doses falling within the intersection range.
150,175	150 and 175 mg/day are the only doses falling within the intersection range.

175,200	175 and 200 mg/day are the only doses falling within the intersection range.

- A further stratification of theoretically responders was introduced according to their baseline characteristics (i.e., $Circ0_{PLT}, Circ0_{WBC}, Circ0_{HCT}$ parameters), as reported in Table S3.2. The condition with normal values of all $Circ0_{PLT}, Circ0_{WBC}$ and $Circ0_{HCT}$ is missing as all PV patients have at least one haematological parameter not in the target range at baseline [45].

Table C.4: Groups defined according to PV patients baseline conditions.

Baseline characteristic within normality range (Yes/No)			**Group**
$Circ0_{PLT}$	$Circ0_{WBC}$	$Circ0_{HCT}$	
No	No	No	C1
Yes	No	No	C2
Yes	Yes	No	C3
No	Yes	No	C4
No	Yes	Yes	C5
No	No	Yes	C6
Yes	No	Yes	C7

- The 7 groups defined in Table C.4 were combined with the 14 categories listed in Table C.3, thus defining 98 subgroups of theoretical respondent virtual patients.
- As illustrated in Figure S3.1, each virtual population was obtained by randomly sampling one individual for each of the 98 subgroups.

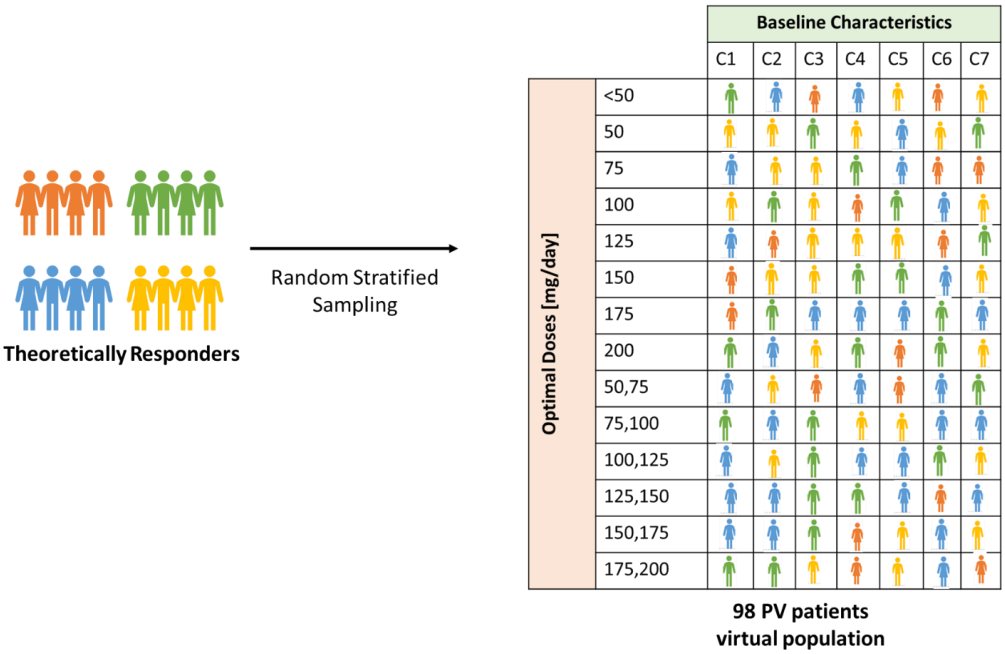


Figure C.5: Stratified random sampling strategy for building the final virtual population.

C.4. QL algorithm hyperparameters

Table C.5: Hyperparameters used to train the QLpop-agent.

Hyperparameter	**Value**
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.99
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{10000}\right)\right)\right)$ with i being the current iteration number

Table C.6: Hyperparameters used to train the QLind-agents.

Hyperparameter	**Value**
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.97
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{10000}\right)\right)\right)$ with i being the current iteration number

C.5. Tuning the reward function to learn a unique adaptive dosing protocol for the whole population with QL

The aim of this section is to provide a comparison of the performances of the QLpop-agent characterized by the first reward (Eqs. 23-33), abbreviated here with Rew1) and by the second one with a smoother penalization on severe toxicities (Eqs.38-40, abbreviated here with Rew2). This assessment was performed on the training virtual population, which is the same in both the cases.

The effect of the two reward functions on the QLpop-agent policies can be seen in Figure C.6. For Rew1 (Panel A) and Rew2 (Panel B), the median (black line) and 90% C.I. (shaded blue area) of the individual profiles in the population are reported. Due to the strong penalties assigned to the severe toxicities in Rew1 (i.e., reward=0, Eq.4), the corresponding QLpop-agent avoided the choice of higher dose levels and totally discouraged the use of the maximum tolerated dose (i.e., 200 mg/day) (see the distribution of each dose levels within the eight months timeframe reported in Figure C.7). Form one hand, this conservative dose strategy prevented a longer permanence of PLT and WBC below $75 \times 10^9/L$ and $3 \times 10^9/L$ (red shaded area in Figure C.6), respectively, from the other hand it was unable to satisfying bring the haematological parameters in the efficacy range.

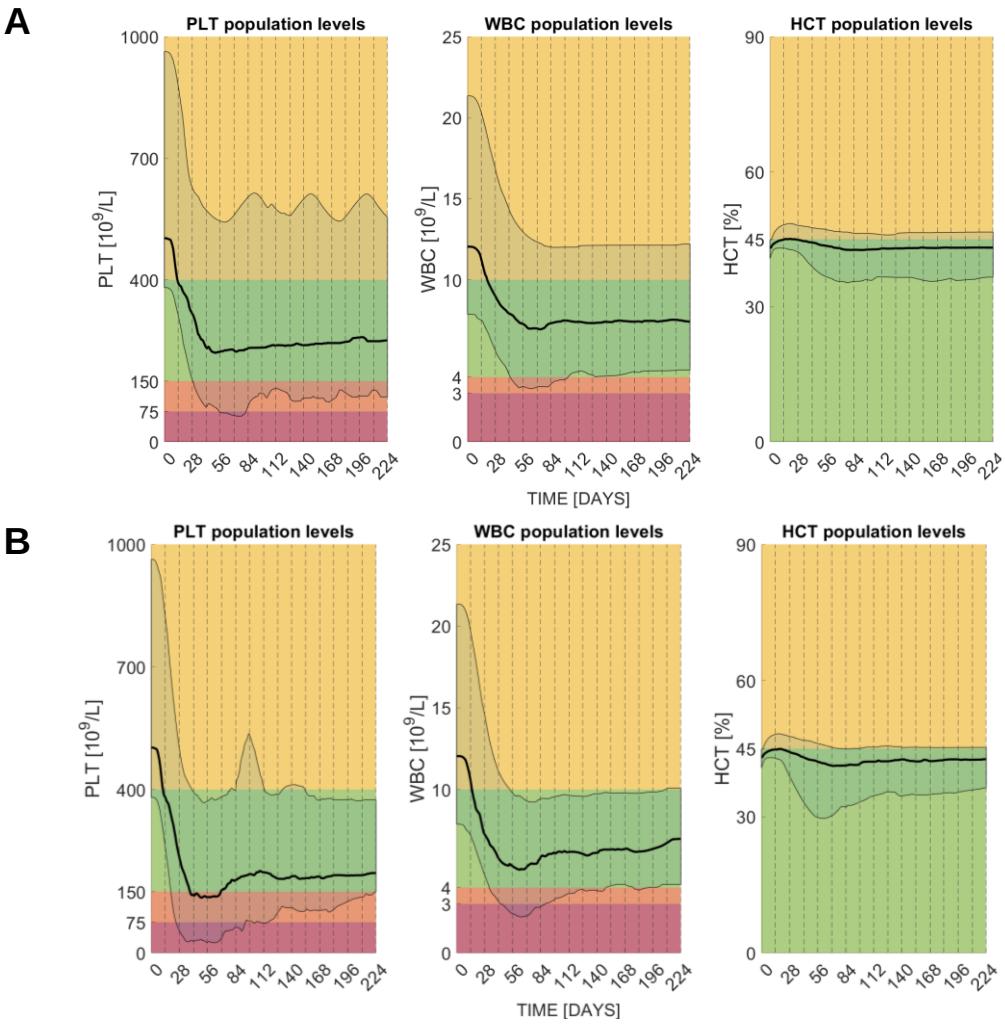


Figure C.6: Effect of Rew1 (Panel A) and Rew2 (Panel B) reward functions on the performances of the QLpop-agent. Results are summarized in terms of median (black line) and 90% C.I. of individual profiles within the population (blue shaded areas). Yellow, green, orange and red shaded areas, represent inefficacy, efficacy, moderate and severe toxicity ranges of each haematological parameter.

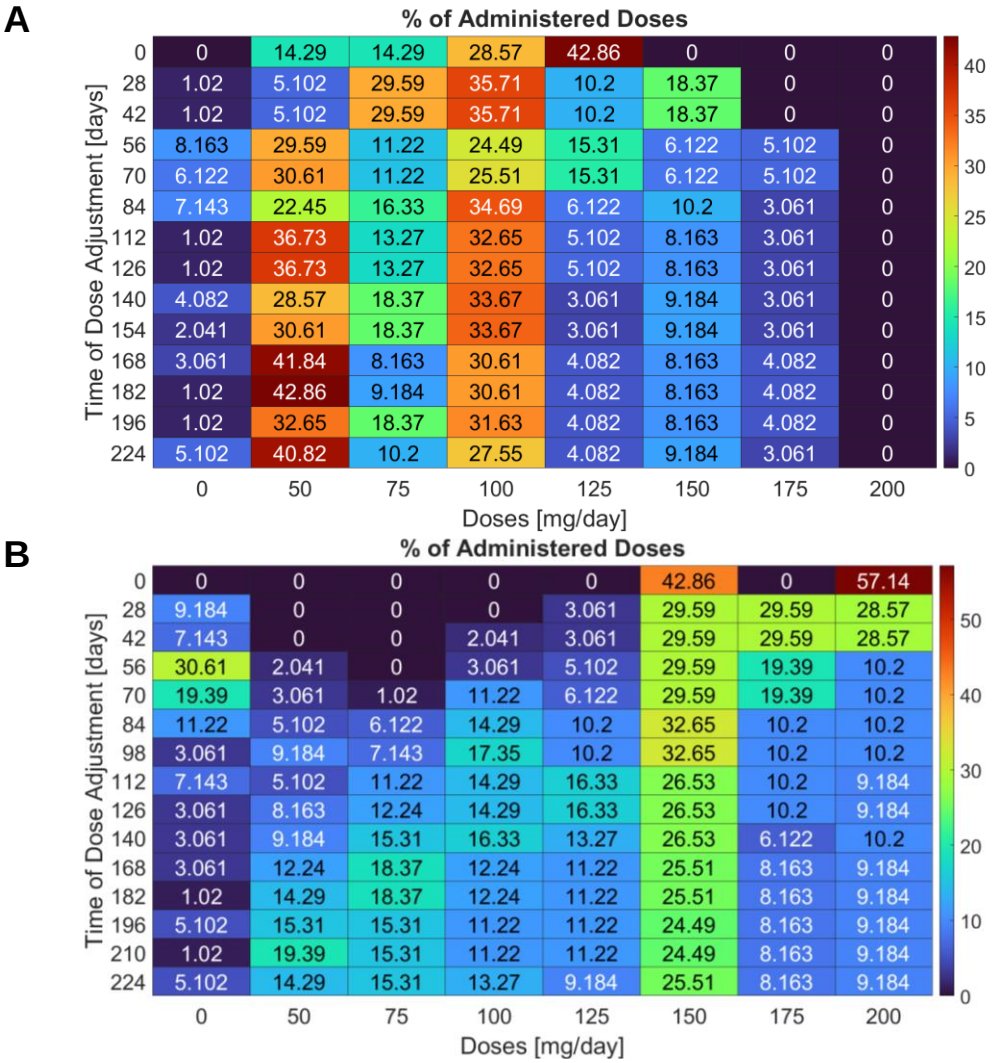


Figure C.7: Distribution of doses administered by the QLpop-agents with Rew1 (Panel A) and Rew2 (Panel B) reward functions for each treatment cycle.

Appendix D

Supplementary Materials of Chapter 5

This appendix contains further information on the application of the individually tailored MIRL approach (section 2.2.2) on the axitinib/anti-hypertensive medication precision dosing problem presented in Chapter 5.

D.1. Empirical Pop-PK-PD-OS model for the co-administration of axitinib and anti-hypertensives

As illustrated in Figure D.1, an empirical PK-PD-OS model was developed to describe the PK-PD processes of AX-AH co-administration and the impact of the anticancer treatment on patient survival probability. This modelling framework was obtained by combining two already published models, one characterizing AX PK-PD-OS and another one AH PD effect on diastolic blood pressure (dBP) [131,157].

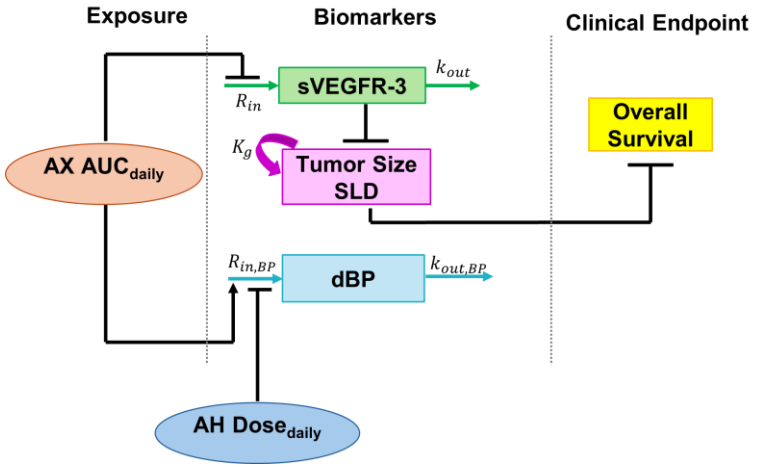


Figure D.1: Schematical representation of the final PK-PD-OS modelling framework used to characterize AX-AH co-administration.

More in details, the original AX PK-PD-OS model directly links the effect of AX on efficacy biomarkers by leveraging drug daily exposure (AUC_{Daily}) due to the very short half-life of this compound [126]. AUC_{Daily} increased dBP levels and lowers the plasma concentration of the soluble version of VEGFR (sVEGFR). The modelling framework in [126] considers only a particular sVEGFR, sVEGFR-3, to describe the anti-angiogenic effect of AX in the tumor, which is measured as the sum of the longest lesions diameters (SLD). Finally, in the OS model, only SLD was found significantly impacting on patient survival probability.

Due to the absence of PK-PD models on AX-AH co-administration, the PD model describing AH effect on dBP during levatinib treatment was integrated in the AX PK-PD-OS framework [131]. This was fundamental to simulate the therapeutic problem optimization in a more realistic scenario, as AH are often administered during AX treatment [130]. Of course, this empirical approach introduces an additional hypothesis not fully validated to the modelling framework. However, this hypothesis is supported by the fact that the introduced AH effect model was estimated on data coming from the co-administration with a compound belonging to TKI family, as Axitinib.

The following subsections will provide a detailed description the components of the adopted modelling framework.

D.1.1. PK model of axitinib daily exposure

For the sake of simplicity, since the driver of AX PD is its daily exposure, in this section only the elements necessary to compute the values and the variability of the AUC_{Daily} will be presented. A comprehensive description of the complete AX PK model can be found in [126].

AX AUC_{Daily} is computed by applying Eq. D1, where F represents the bioavailability. The PK model assumes that CL is lognormally distributed

and part of its IIV is described by the following categorical covariates: Age>60/Age ≤ 60, Race Japanese/Not Japanese, Active Smoke Status/Inactive Smoke Status. Table D.1 reports the estimated AX PK parameters with their IIV in the population.

$$\begin{cases} AUC_{daily} = \dfrac{Dose_{Daily} \cdot F}{CL} \\ CL_{typical} = CL_{pop} \cdot \left(1 - \beta_{Age} \cdot Age_{60}\right) \cdot \\ \cdot \left(1 - \beta_{Race} \cdot Race_{Japanese}\right) \cdot \left(1 + \beta_{smoker} \cdot Smkoe_{Active}\right) \end{cases}$$

(D. 1)

Table D.1: Parameter values of AX PK model. CL IIV is expressed as CV of the lognormal distribution.

Parameter [units]	Estimated Value
CL_{pop}[L/h]	14.6
F	0.402
β_{Age}	0.213
β_{Race}	0.249
β_{smoker}	1.02
CV of IIV on CL [%]	59.9

D.1.2. PK-PD model of axtinib effect of dBP

In [157], AX effect on dBP is described through an E_{max} model which is driven by AUC_{daily} (Eq. D.2).

$$\begin{cases} \dfrac{ddBP}{dt} = R_{in.dBP} \cdot \left(1 + \dfrac{E_{max} \cdot S_0 \cdot AUC_{daily}}{E_{max} + S_0 \cdot AUC_{daily}}\right) - k_{out,dBP} \cdot dBP \\ k_{out,dBP} = \dfrac{1}{MRT}, R_{in,dBP} = k_{out,dBP} \cdot Base_0, S_0 = E_{max}/AU\,C_{50} \\ dBP(0) = Base_0 \end{cases}$$

(D. 2)

In the original work, PK-PD model parameters of dBP were estimated on a population of Japanese individuals which were normotensive at baseline, accordingly to treatment eligibility criteria. Moreover, in the model building and parameter estimations, only data of the first treatment cycle were used to exclude patients taking anti-hypertensive to control AX-induced hypertension. IIV was applied only on the baseline dBP, $Base_0$, which follows a logormal distribution. To describe the skewed distribution of $Base_0$, random effects were remapped trough a Box Cox transformation with shape parameter ϕ. Table D.2 summarizes the PK-PD model parameter values of dBP.

Table D.2: PK-PD model parameter values of dBP.

Parameter [units]	Estimated Value
PD parameters	
$Base_0$[mmHg]	78.9
ϕ[-]	-5.42
MRT [days]	4.92
$E_{max}[-]$	0.197
$S_0[L \cdot h^{-1} \cdot \mu g^{-1}]$	0.00127
CV of IIV on $Base_0$ [%]	6.7

D.1.3. Empirical AX-AH PK-PD model of dBP

The modelling framework introduced in the previous section does not describe the effect of AH medications administered to compensate for AX-induced hypertension. Therefore, a PK-PD model describing the co-administration of a TKI compound (i.e., levatinib) with AHs was integrated within AX PK-PD model of dBP by modifying Eq. D.2 as follows:

$$\begin{cases} \dfrac{ddBP}{dt} = R_{in,dBP} \cdot \left(1 + \dfrac{E_{max} \cdot S_0 \cdot AUC}{E_{max} + S_0 \cdot AUC}\right) \cdot \dfrac{1}{1 + \theta_{AH} \cdot DDE} - k_{out,dBP} \cdot dBP \\ DDE = \sum_i^N \dfrac{DD_i}{SDD_i} \\ k_{out,dBP} = \dfrac{1}{MRT}, R_{in} = k_{out,dBP} \cdot Base_0, S_0 = E_{max}/AU\,C_{50} \\ dBP(0) = Base_0. \end{cases}$$

(D. 3)

In particular, DDE is the daily dose equivalent which represents the cumulative daily dose of all N AH drugs taken by the patient. For the *i-th* AH, the daily amount, DD_i, is normalized by its standard daily dose, SDD_i, defined by WHO [158]. θ_{AH}, representing the AH lowering effect on dBP, was fixed to 0.036 [-], which is the value estimated in [131].

Thus, replacing Eq.D.2 with Eq. D.3 allows to simulate the effect on dBP of AX- AH co-administration.

D.1.4. PK-PD model of axitinib effect on sVEGFR-3

Similarly to dBP, an indirect response PK-PD model was adopted also for AX effect on sVEGFR-3 [157]. In particular, the AX AUC_{Daily} decreases the production rate of sVEGFR-3 concentration, $R_{in,sVEGFR3}$, with a saturation of the inhibition effect (Eq. D.4):

$$\begin{cases} \dfrac{dsVEGFR3}{dt} = R_{in,sVEGFR3} \cdot \left(1 - \dfrac{1}{AUC_{50,\text{sVEGFR3}} + AUC_{Daily}}\right) - k_{out,sVEGFR3} \cdot sVEGFR3 \\ k_{out,sVEGFR3} = \dfrac{1}{MRT}, \qquad R_{in,sVEGFR3} = k_{out,sVEGFR3} \cdot sVEGFR3_0, \\ sVEGFR3(0) = sVEGFR3_0 \end{cases}$$

(D. 4)

Table D.3 reports the parameter values of the PK-PD model on sVEGFR-3. IIV was assumed to be lognormally distributed.

Table D.3: Parameter values of PK-PD model for sVEGFR3

Parameter [units]	Estimated Value
$sVEGFR3_0$[pg/mL]	78.9
$AUC_{50,\text{sVEGFR3}}$ $[\mu g \cdot h/L]$	717
MRT [days]	5.76
CV of IIV on $sVEGFR3_0$[%]	49
CV of IIV on AUC_{50} [%]	45

D.1.5. PK-PD model of axitinib effect on SLD

In the modelling framework developed in [157], AX indirectly affect tumor growth (Figure D.1). As reported in Eq. D.5, the sum of the longest SLD is directly linked to the concentration levels of the sVEGFR3. In particular, the relative decrease of sVEGFR3 with respect to its baseline value ($sVEGFR3_0$), $sVEGFR3_{rel}$ (Eq. D.6), is used in the model to describe the tumor growth inhibition. This effect is time varying as it follows an exponential decay governed by λ .

$$\begin{cases} \dfrac{dSLD}{dt} = k_g \cdot SLD - k_{sVEGFR3} \cdot sVEGFR3_{rel} \cdot e^{-\lambda \cdot t} \cdot SLD \\ SLD(0) = SLD_0 \end{cases}$$

(D. 5)

$$sVEGFR3_{rel} = \frac{sVEGFR3_0 - sVEGFR3}{sVEGFR3_0}$$

(D. 6)

Table D.4 summarizes the estimated values of PK-PD model parameters for SLD. IIV is lognormally modelled also in that case. In the original model, the initial tumor size, SLD_0, was fixed to the baseline observation [157].

Table D.4: Values of PK-PD model parameters for SLD.

Parameter [units]	Estimated Value
$k_g[week^{-1}]$	0.00361
$k_{sVEGFR}[week^{-1}]$	0.176
$\lambda[week^{-1}]$	0.101
CV of IIV on k_g [%]	160
CV of IIV on λ[%]	72

D.1.6. PK-PD-OS model of axitinib

Hazard rate of death was modelled leveraging a loglogistic function for baseline risk. SLD effect was introduced in the model by considering a multiplicative term exponentiating the product $\beta_{SLD} \cdot SLD$ (Eq. D.7).

$$h(t) = \frac{\psi^{\frac{1}{\gamma}} \cdot t^{\frac{1}{\gamma}-1}}{\gamma \cdot \left(1 + (\psi \cdot t)^{\frac{1}{\gamma}}\right)} \cdot e^{\beta_{SLD} \cdot SLD}, \qquad \psi = e^{-\beta_0}$$

(D. 7)

In particular, β_{SLD} represents the increase in the hazard rate of death due to 1mm increase of SLD. Table D.5 reports the values of the OS model parameters.

Table D.5: Parameter values for the OS model.

Parameter [units]	Estimated Value
$\beta_0[-]$	7.09
γ[-]	0.298
$\beta_{SLD}[mm^{-1}]$	0.0115

D.1.7. Assumption on AH medications following steady-state analysis of dBP PK-PD model

A mathematical analysis of the dBP modelling framework was performed considering the steady state. From Eq. D.3, it is possible to derive the expression of dBP at steady state, dBP_{SS}, as a function of AX AUC_{Daily}:

$$dBP_{SS} = Base_0 \cdot \left(1 + \frac{E_{max} \cdot AUC_{Daily}}{AUC_{50} + AUC_{Daily}}\right) \frac{1}{1 + \theta_{AH} \cdot DDE}.$$

(D. 8)

Considering AX monotherapy (i.e., AH=0 in Eq. D.8), the effect of the anticancer treatment on dBP_{SS}, $dBP_{Base\wedge AX}$, is described by the following E_{max} model (Eq. D.9):

$$dBP_{Base\wedge AX} = Base_0 \cdot \left(1 + \frac{E_{max} \cdot AUC_{Daily}}{AUC_{50} + AUC_{Daily}}\right).$$

(D. 9)

Following the eligibility criteria in [117,125,127,128], all patients are normotensive before starting treatment (i.e., $Base_0 < 90mmHg$). Furthermore, under the assumption of reaching AX E_{max} effect, by replacing the values of the parameters in Eq. D.9, $dBP_{Base\wedge AX}$ will be $< 107.7mmHg$. Therefore, according to this computation, it is impossible to observe, in these patients, a dBP higher than $107.7mmHg$ during AX treatment. It is reasonable that the effect of AH medications on $dBP_{Base\wedge AX}, AH_{effect}$ (Eq. D.10), allows to counteract the maximum AX-induced increase by bringing dBP within the target range of 90 to 100 mmHg.

Consequently, the available AH daily dose equivalent (DDE) levels were set to {0, 1, 2, 3, 4}. Indeed, by replacing DDE values in Eq. D.10, AH medications allow 4-13% reduction in $dBP_{Base\wedge AX}$, which is sufficient to bring dBP back into the [90,100] mmHg range, even in the worst-case scenario (i.e., when $dBP_{Base\wedge AX}$= 107.7 mmHg).

$$AH_{effect} = \frac{1}{1 + \theta_{AH} \cdot DDE}$$

(D. 10)

D.1.8. Steady-state analysis of SLD PK-PD model

Since the SLD model directly depends on sVEGFR3 levels (Eq. D.5 and D.6), it is first necessary to derive the steady-state level of sVEGFR3, $sVEGFR3_{ss}$. By zeroing Eq. D.4, the expression of $sVEGFR3_{ss}$ is obtained (Eq. D.11).

$$sVEGFR3_{SS} = sVEGFR3_0 \cdot \left(1 - \frac{AUC_{Daily}}{AUC_{50,\text{sVEGFR3}} + AUC_{Daily}}\right)$$

(D. 11)

By replacing the expression in Eq. D.11 in Eq. D.6, the relative shift from $sVEGFR3_0$ at steady state, $sVEGFR3_{rel.SS}$, is obtained (Eq. D.12).

$$sVEGFR3_{rel.SS} = \frac{AUC_{Daily}}{AUC_{50,\text{sVEGFR3}} + AUC_{Daily}}$$

(D. 12)

Then, by setting Eq. D.6. equal to zero and replacing $sVEGFR3_{rel}$ $sVEGFR3_{rel.SS}$, Eq. D.13 is obtained.

$$SLD_{SS} \cdot \left(K_g - K_s \cdot \frac{AUC_{Daily}}{AUC_{50,\mathrm{sVEFR3}} + AUC_{Daily}} \cdot \mathrm{e}^{-\lambda \cdot \mathrm{t}} \right) = 0$$

(D. 13)

In particular, the time at which AX is no longer effective, t_{ineff}, can be obtained by solving Eq. 13 with respect to t and using the $AUC_{Daily|AX=10}$ i.e., the AUC_{Daily} following an AX administration of 10 mg b.i.d. (Eq. 14).

$$t_{ineff} = -\frac{1}{\lambda} \cdot \ln\left(\frac{k_g}{k_{sVEGFR3}} \cdot \frac{AUC_{50,sVEGFR3} + AUC_{Daily}}{AUC_{Daily}} \right)$$

(D. 14)

D.2. Generation of the virtual test population

D.2.1. Statistical distributions of patient covariates

Virtual patients covariates were described by statistical distributions whose parameters were fixed considering the population used in [126,157] to develop the AX PK-PD-OS modelling framework. This strategy allows to generate a virtual test population coherent with the real patients used to estimate model parameters.

As reported in Eq. D.1, covariates were introduced in the model to explain part of AX CL IIV. More specifically, Age, Japanese Race and Smoker status were included in the final PK model [126]. Since the AX PK-PD-OS model was originally identified only on Japanese patients, this covariate was fixed in the virtual population.

Smoker status was extracted from a binomial distribution with $p(Active\ Smoke\ Status) = 0.5$ as information on its distribution were not reported in AX literature works [125–128].

In the dataset on which the AX PK-PD-OS model was identified, median age was 63 years, and its range was [34,80] years. Therefore, it was assumed that $Age \sim N(63,7.24)$ with standard deviation approximated by $(median - min)/4$. Moreover, the underlying standardized normal distribution of age was truncated between [-4,2.34]. As illustrated in Figure D.2, this strategy allowed to extract age samples falling in the [34,80] years range with a median of 63 years, approximately.

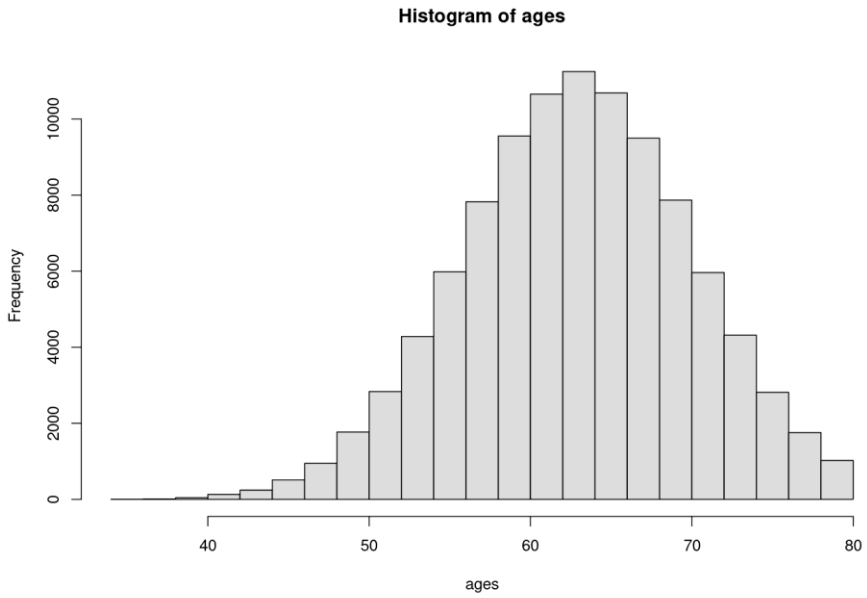


Figure D.2: Histogram of the distribution of ages in the virtual population. Samples of age were generated by sampling from a normal distribution with parameters $\mu = 63$ [years] and $\sigma = 7.25[years]$.

In the original AX PD-PD-OS model (section D.1.5) the measured tumor sizes were used as the initial point of SLD dynamics (SLD_0). However, the distribution of SLD_0 was not reported in [157]. Therefore, it was assumed lognormally distributed as described in Eq. after a qualitative extrapolation from the plots in [157].

$$SLD_0 = 100 \cdot e^{\omega}, \text{ with } \omega \sim N(0, 0.5^2)$$

(D. 15)

D.2.2. Stratified random sampling to generate the virtual patient population

By considering all sources of IIV on both parameters and covariates (section D.2.1 and tables D.1-4), a virtual population of 100,000 patients was generated. Then, Eq. D.9 was applied to compute the dBP at steady state for each AX dosage (i.e., 2,3,5,7,10 mg b.i.d.) and each subject in the virtual population. The analysis led to identify the following groups based on patients dBP levels at steady state:

- **Group 1:** 39.18 % of patients cannot reach the target dBP range [90,100) mmHg, even with the highest AX dose (10mg b.i.d.). According to their baseline dBP values, patients were split into three subgroups (**Group 1.1** if baseline dBP<70 mmHg, **Group 1.2** if

baseline dBP in [70,80) mmHg range, **Group 1.3** if baseline dBP in [80,90) mmHg range);

- **Group 2:** 55.6% of patients can reach the target dBP range [90,100) mmHg when treated with the highest AX dose (10 mg b.i.d.). According to their baseline dBP values, patients were split into three subgroups (**Group 2.1** if baseline dBP in [70,80) mmHg range, **Group 2.2** if baseline dBP in [80,90) mmHg range);
- 8.35% of patients can reach the target dBP range [90,100) mmHg when treated with doses lower than 10 mg b.i.d. However, when they are treated with all the higher doses, moderate toxicities occur (dBP range [100,105] mmHg):
    - **Group 3.1:** 0.5%, 2 mg b.i.d.
    - **Group 3.2:** 1.32% 3 mg b.i.d.
    - **Group 3.3:** 1.31% 5 mg b.i.d.
    - **Group 3.4:** 1.74% 7 mg b.i.d.
- 0.075% of patients has at least one AX dose that can cause a dangerous hypertension condition (> 105 mmHg). In particular:
    - **Group 4.1:** 1 subject in which that happens starting from 2 mg b.i.d. of AX. Therefore, all the doses lead to severe hypertension.
    - **Group 4.2:** 6 subjects in which that happens starting from a 3 mg b.i.d. of AX. None of them can reach the target range with lower doses.
    - **Group 4.3:** 6 subjects in which that happens starting from a 5 mg b.i.d. of AX. None of them can reach the target range with lower doses.
    - **Group 4.4:** 14 subjects in which that happens starting from a 7 mg b.i.d. of AX. None of them can reach the target range with lower doses.
    - **Group 4.5:** 41 subjects in which that happens at 10 mg b.i.d. of AX. None of them can reach the target range with lower doses.
    - **Group 4.6:** 7 subjects in which that happens at 10 mg b.i.d. of AX. However, a dose of 2 mg b.i.d. is sufficient to reach the target range.
- **Group 5:** 0.24% of patients will experiment a moderate hypertension (range [100,105] mmHg) for all the AX doses.

Then, for each patient in the 16 dBP response groups, Eq. D.14 was applied to compute the time from which AX loses its efficacy. These values were stratified in 5 scenarios depending on the cycle at which AX becomes ineffective. Therefore, as reported in Table D.6, patients were stratified according to both their dBP and tumor resistance patterns.

Table D.6: Percentages of patients in the groups obtained from the stratifications based on dBP response and time of complete loss of AX efficacy.

dBP response Group	Time at which AX is no longer effective				
	< 6 cycles	6-12 cycles	12-18 cycles	18-24 cycles	>24 cycles
Group 1.1	42.71	29.76	13.33	6.26	8.94
Group 1.2	44.35	29.08	12.94	6.10	7.53
Group 1.3	55.01	25.01	10.42	4.40	5.16
Group 2.1	44.05	30.27	14.24	6.93	4.51
Group 2.2	43.24	29.26	13.32	6.29	7.89
Group 3.1	37.84	28.89	15.07	8.23	9.97
Group 3.2	38.72	29.83	14.75	7.43	9.27
Group 3.3	39.09	30.21	14.61	7.02	9.07
Group 3.4	39.86	30.47	14.18	6.71	8.78
Group 4.1	100	0	0	0	0
Group 4.2	50	12.5	25	12.25	0
Group 4.3	28.57	33.15	21.43	3.57	13.28
Group 4.4	41.05	29.47	9.47	10.52	9.49
Group 4.5	38.78	32.25	11.52	10.59	6.86
Group 4.6	44.44	18.52	11.11	18.51	7.42
Group 5	35.83	30.53	15.55	7.60	10.49

The virtual population was created by randomly sampling one patient from each group listed in Table D.6. Since some tumor-dynamics patterns did not occur in some dBP response group, the virtual population consisted of 75 patients instead of 80 (i.e., one for all the possible combinations).

D.3. Hyperparameters of QL algorithm

Table D.7 reports QL hyperparameters for the AX-AH co-administration precision dosing problem. For two virtual patients, when the S&LT-Reward function was used for the QLind-agents training, the optimal learning rate was $\alpha = 0.05$ and the number of iterations was fixed to 75000.

Table D.7: Hyperparameters for the QL algorithm in the AX-AHs co-administration.

Hyperparameter	Value
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.95
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{25000}\right)\right)\right)$ with i being the current iteration number

D.4. Supplementary figures for QLind-agents trained with S&LT-Reward function

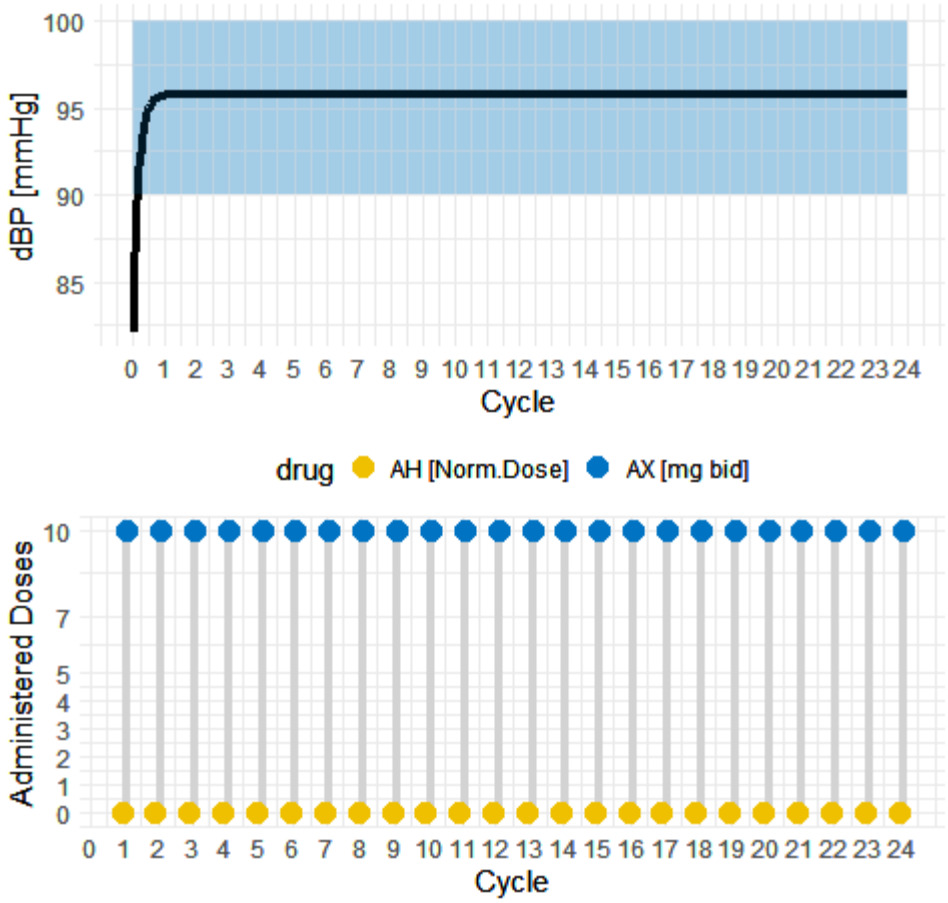


Figure D.3: Example in which QLind-agent trained with S&LT-Reward is able to detect patient well tolerating the highest AX dose since the beginning of treatment

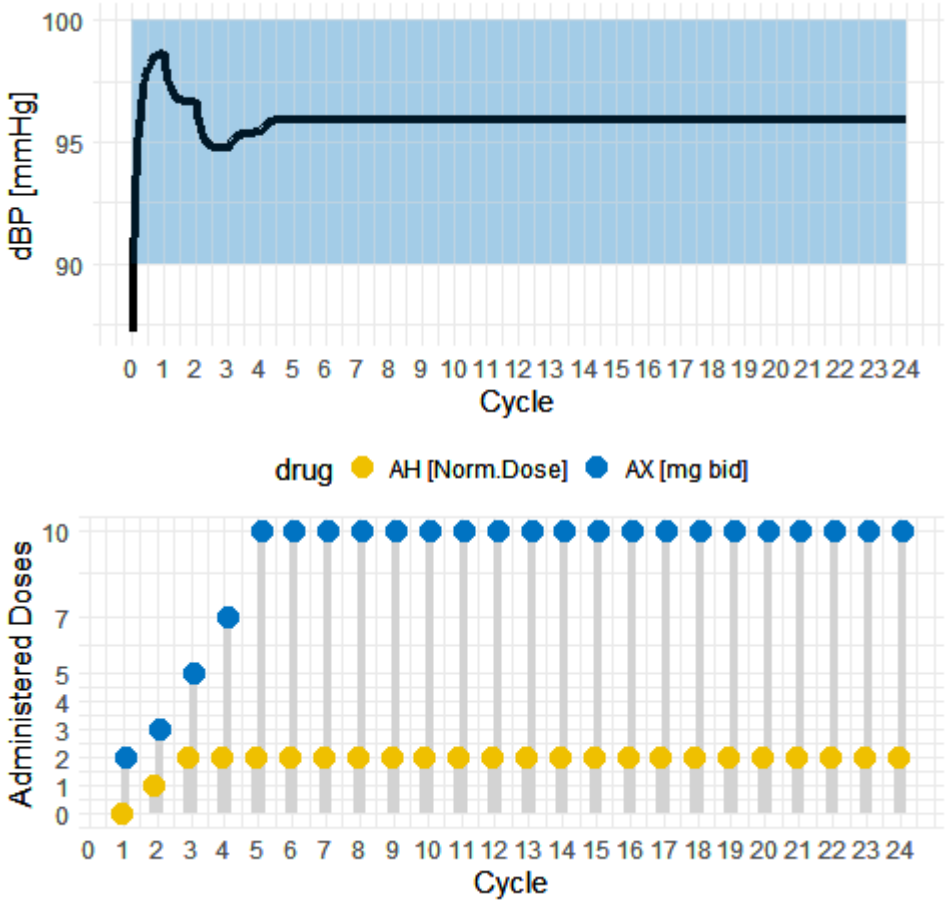


Figure D.4: Example of QLind-agent trained with ST-Reward personalizing AX-AH co-administration with a joint up-titration strategy.

Appendix E

Supplementary Materials of Chapter 6

This appendix contains further information on the extensions of the MIRL approaches presented in Chapter 6.

E.1. Simplified version of givinostat precision dosing problem

This section provides a detailed description of the simplified givinostat precision dosing problem applied to investigate the performances of the Bayesian MIRL approach presented in section 6.1.

As concerns givinostat clinical setting described in section 4.1.1 , the available drug dose levels were reduced from {50, 75, 100, 125, 150, 175, 200} to {50, 100, 150, 200} mg/day, coherently to the original design of the early clinical studies on PV patients [45].

The other changes are related to the formalization of givinostat precision dosing problem as MDP. As done before, patient health status was described by combining a discrete representation of the observed PLT, WBC and HCT values with the information on the previous administered dose (PrevDose). However, the reduction of the available givinostat doses led to change the $PrevDose$ information within patient state by replacing Eq. 37 with Eq. E1:

$$PrevDose = \{0, 50, 100, 150, 200\}\ \text{mg/day}\ (for\ all\ treatment\ stages). \quad (E.1)$$

In addition, only a single toxicity range, respectively for thrombocytopenia and neutropenia, was adopted. Thus, Eqs. 34 and 35 were replaced by Eqs. E2 and E3. Differently, HCT was discretized as in Eq. 36.

$$PLT_{Discr}(PLT_{Obs}) = \begin{cases} 1 & \text{if } PLT_{Obs} < 150 \times 10^9/L \text{ (Thrombocytopenia)} \\ 2 & \text{if } PLT_{Obs} \in [150,400] \times 10^9/L \ (Efficacy) \\ 3 & \text{if } PLT_{Obs} > 400 \times 10^9/L \ (Unefficacy) \end{cases}$$

(E. 2)

$$WBC_{Discr}(WBC_{Obs}) = \begin{cases} 1 & \text{if } WBC_{Obs} < 4 \times 10^9/L \ (Neutropenia) \\ 2 & \text{if } WBC_{Obs} \in [4,10] \times 10^9/L \ (Efficacy) \\ 3 & \text{if } WBC_{Obs} > 10 \times 10^9/L \ (Unefficacy) \end{cases}$$

(E. 3)

As concerns action selection, QL-agents were constrained to perform stepwise dose changes (i.e. increase/decrease by one level, $D+/D-$) or to maintain the current dose level ($D=$) independently by the observed PLT, WBC and HCT values. Therefore, the onset of severe/moderate toxicities did not trigger by default the temporary treatment interruption. Indeed, QL-agents could decide to stop givinostat treatment for at least one cycle (28 days) following a 50 mg/day administration (lowest possible administrable drug amount. After the temporary interruption, QL decides whether to resume the treatment with 50 mg/day (fixed) or not. Analogously to the formalization adopted in Chapter 3, also in this case QL-agents could select the initial dose for the treatment. Table E.1 summarizes all the constraints adopted on QL-based action selection.

Table E.1: Summary of constraints introduced on the QL-based action selection for the simplified version of givinostat precision dosing problem.

$PrevDose$	**QL-Agent Actions**
Initial State	
-1	$50, 100, 150, 200\ [mg/day]$
For each value of PLT_{Discr}, WBC_{Discr} and HCT_{Discr}	
0	D=, D+
50	$D=, D+, D-$
100	$D=, D+, D-$
150	$D=, D+, D-$
200	$D=, D-$

The reward function described in section 4.2.2 was used to train QLind, QLind-bay and QLpop agents.

E.2. Generation of the virtual population

The stratified random sampling of the virtual patient population used to evaluate the Bayesian MIRL approach and to train the QLpop-agent was performed following the procedure detailed in section C.3 of Appendix C. In

this case, the only difference lies in the definition of patient groups according to givinostat theoretical optimal dose.

Due to the reduction of the available givinostat doses (section E.1) a smaller number of response patterns was considered. Specifically, the groups originally defined by the intersections of the optimal dose ranges for PLT, WBC, and HCT in Table C.3 were replaced by those in Table E 2.

Table E 2: Groups defined according to givinostat theoretical optimal dose.

Group	Description
<50	The upper bound of the intersection range is below 50 mg/day.
50	50 mg/day is the only dose falling within the intersection range.
100	100 mg/day is the only dose falling within the intersection range.
150	150 mg/day is the only dose falling within the intersection range.
200	200 mg/day is the only dose falling within the intersection range.
50,100	50 and 100 mg/day are the only doses falling within the intersection range.
100,150	100 and 150 mg/day are the only doses falling within the intersection range.
150,200	150 and 200 mg/day are the only doses falling within the intersection range.

Therefore, by combining these groups with the seven in Table C.4, which are defined according to baseline PLT, WBC, and HCT levels, 56 givinostat response patterns were identified in this case. For the training set of the virtual population, one patient was randomly sampled for each of these groups. Differently, the virtual population on which the Bayesian MIRL approach was evaluated was generated by randomly extracting two virtual patients from each of the 56 subgroups.

E.3. QLpop-agent used in the Bayesian MIRL approach

A QLpop-agent was trained on a heterogeneous pool of 56 virtual PV patients (see section E.2 for details on its extraction) considering the formalization of givinostat precision dosing problem in section E.1 and the treatment duration of 8 months, i.e., the average time to achieve a stable complete haematological response [45]. Table E.3 and Table E.4 report

algorithm hyperparameters and the QLpop-agent performances on the training population, respectively. To evaluate the generalizability of the QLpop-based dosing protocol, the QL-based dosing strategy was applied and benchmarked was and benchmarked by the clinical protocol on the 112-patients virtual population used for the Bayesian MIRL. From the results in Table E.3, it emerges that QLpop-agent reached similar efficacy performances to the clinical protocol, thus confirming a with the good generalizability of the learnt dosing protocol.

Table E.3: Hyperparameters of QL algorithm used to train QLpop-agent.

Hyperparameter	**Value**
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.99
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{25000}\right)\right)\right)$ with i being the current iteration number

Table E.4: Comparison between the QLpop-agent and givinostat clinical protocol on the training virtual population.

QLpop-agent	**Clinical protocol**
PLT response rate at 8th month [%]	
96.4	91.2%
WBC response rate at 8th month [%]	
96.4	91.2%
HCT response rate at 8th month [%]	
92.8	89.5%
CHR at 8th month [%]	
86.0	75.0%
Median time needed to achieve a first CHR [days] (standard deviation)	
40.5 (28.98)	47.5 (27.45)

Table E.5: Comparison between the QLpop-agent and givinostat clinical protocol on the 112-patients virtual test population

QLpop-agent	**Clinical protocol**
PLT response rate at 8th month [%]	
88.39	90.1%
WBC response rate at 8th month [%]	
91.0	90.1%
HCT response rate at 8th month [%]	

93.7	87.5%
CHR at 8th month [%]	
74.7	74.1%
Median time needed to achieve a first CHR [days] (standard deviation)	
47 (26.1)	51 (26.7)

E.4. Hyperparameters of QLind and QLind-bay agents

Table E.6: Hyperparameters used to train QLind and QLind-bay agents.

Hyperparameter	**Value**
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.97
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{10000}\right)\right)\right)$ with i being the current iteration number

E.5. Prioritized Sweeping QL

The original formulation of Prioritized Sweeping QL proposed for a deterministic system is reported in Algorithm E.1 [76]. Its extension to a stochastic system can be obtained by using Eq. 55 instead of the canonical QL update formula (Eq. 14). The attention to be given at a certain s, a couple is regulated by the threshold ψ which is a model hyperparameter.

Algorithm E.1: Pseudocode of Prioritized Sweeping QL algorithm

Given: set of N states, set of M actions, learning rate α, discount factor γ, a probability ϵ, a maximum number of training iterations I, selective memory E, transition memory D, attention threshold ψ
init Q matrix arbitrarily, empty E
loop for each episode (I times):
    Set the current system state to S_0
    loop for each decisional time step t:
        p ← uniform random number ∈ [0,1]
        if p<ϵ
            Select action A_t randomly
        else
            $A_t \leftarrow arg\,max_a\, Q(S_t, a)$

Perform A_t on the system
Observe next state S_{t+1} and reward R_{t+1}
Store if absent $< S_t, A_t, S_{t+1}, R_{t+1} >$ in D
$U = |R_{t+1} + \gamma \cdot max_a\, Q(S_{t+1}, a) - Q(S_t, A_t)|$
if $U > \psi$
descending storage of S_t, A_t in E according to U
loop while E is not empty (K times):
$S_t^k, A_t^k \leftarrow first(E)$
Perform A_t^k when the system is in S_t^k
Observe next state S_{t+1}^k and reward R_{t+1}^k
$Q(S_t^k, A_t^k) =$
$Q(S_t^k, A_t^k) + \alpha \cdot [R_{t+1}^k + \gamma \cdot max_a\, Q(S_{t+1}^k, a) - Q(S_t^k, A_t^k)]$
Remove S_t^k, A_t^k from E
loop for each S_t^j, A_t^j in D leading to S_t^k:
Get the reward R_{t+1}^j for S_t^j, A_t^j from D
$U = \left|R_{t+1}^j + \gamma \cdot max_a\, Q\left(S_t^k, a\right) - Q\left(S_t, A_t^j\right)\right|$
if $U > \psi$
descending storage of S_t^j, A_t^j in E according to U
Set current system state to S_{t+1}

E.5. Vancomycin Pop-PK model

In this section, a description of the vancomycin Pop-PK model published in [137] and integrated in the EQL framework (sections 6.2.1 and 6.2.3) is presented. A one compartment model was used to describe vancomycin PK in a population of ICU patients having a continuous infusion regimen. IIV was described with a log-normal distribution, body weight and urinary creatinine clearance (CrCL) were included in the model to describe the variation of volume distribution and clearance, respectively (Eqs. E.4 and E.5).

$$V = \theta_{Vol} \cdot Weight \cdot \exp(\eta_V)\text{, with } \eta_V \sim N(0, \omega_V^2)$$

(E. 4)

$$CL = \theta_{CL} \cdot \left({}^{CrCL}/_{100}\right) \cdot \exp(\eta_{CL})\text{, with } \eta_{CL} \sim N(0, \omega_{CL}^2).$$

(E. 5)

The observed vancomycin values, z, are linked to model prediction, y, through a RUV model combining an additive and a proportional term (Eq. E.6).

$$z = y \cdot (1 + \epsilon_1) + \epsilon_2\text{, with } \epsilon_1 \sim N(0, \sigma_1^2) \text{ and } \epsilon_2 \sim N(0, \sigma_2^2)$$

(E. 6)

The parameter values of vancomycin Pop-PK model are reported in Table E.7.

Table E.7: Parameter values of vancomycin Pop-PK model.

Parameter	**Units**	**Value**
θ_{CL}	L/h	4.58
θ_{Vol}	L/Kg	1.53
ω_V	-	0.374
ω_{CL}	-	0.389
σ_1^2	-	0.199
σ_2^2	mg/L	2.4

E.6. Algorithm Hyperparameters

Table E. 8: Training hyperparameters of QLdet-agent.

Hyperparameter	**Value**
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.999
Probability of ϵ -greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{25000}\right)\right)\right)$ with i being the current iteration number

Table E. 9: Training hyperparameters of QLc-agent.

Hyperparameter	**Value**
Number of iterations	50000
Learning Rate	0.1
Discount Factor (γ)	0.999
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{25000}\right)\right)\right)$ with i being the current iteration number

Table E. 10: Training hyperparameters of EQL-agent

Hyperparameter	**Value**
Number of iterations	50000
Discount Factor (γ)	0.999
Probability of ϵ-greedy strategy	$\epsilon = \max\left(0.3, \exp\left(i \cdot \left(-\frac{\ln 0.3}{25000}\right)\right)\right)$ with i being the current iteration number

References

1. Konstantinidou MK, Karaglani M, Panagopoulou M *et al.* Are the Origins of Precision Medicine Found in the Corpus Hippocraticum? *Molecular Diagnosis & Therapy* 2017;**21**:601–6.

2. Sykiotis GP, Kalliolias GD, Papavassiliou AG. Pharmacogenetic principles in the Hippocratic writings. *J Clin Pharmacol* 2005;**45**:1218–20.

3. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53.

4. Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology* 2001;**19**:491–6.

5. Hamburg Margaret A., Collins Francis S. The Path to Personalized Medicine. *New England Journal of Medicine* **363**:301–4.

6. Visvikis-Siest S, Theodoridou D, Kontoe M-S *et al.* Milestones in Personalized Medicine: From the Ancient Time to Nowadays—the Provocation of COVID-19. *Frontiers in Genetics* 2020;**11**.

7. Delpierre C, Lefèvre T. Precision and personalized medicine: What their current definition says and silences about the model of health they promote. Implication for the development of personalized health. *Frontiers in Sociology* 2023;**8**.

8. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington (DC), 2011.

9. Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. *Health Aff (Millwood)* 2018;**37**:694–701.

10. Reardon S. Obama to seek $215 million for precision-medicine plan. *Nature* 2015, DOI: 10.1038/nature.2015.16824.

11. Ashley EA. Towards precision medicine. *Nature Reviews Genetics* 2016;**17**:507–22.

12. Naithani N, Sinha S, Misra P *et al.* Precision medicine: Concept and tools. *Medical Journal Armed Forces India* 2021;**77**:249–57.

13. Wang RC, Wang Z. Precision Medicine: Disease Subtyping and Tailored Treatment. *Cancers* 2023;**15**, DOI: 10.3390/cancers15153837.

14. Muharremi G, Meçani R, Muka T. The Buzz Surrounding Precision Medicine: The Imperative of Incorporating It into Evidence-Based Medical Practice. *Journal of Personalized Medicine* 2024;**14**, DOI: 10.3390/jpm14010053.

15. Johnson KB, Wei W-Q, Weeraratne D *et al.* Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical and Translational Science* 2021;**14**:86–93.

16. Schork NJ. Artificial Intelligence and Personalized Medicine. *Cancer Treat Res* 2019;**178**:265–83.

17. Mesko B. The role of artificial intelligence in precision medicine. *Expert Review of Precision Medicine and Drug Development* 2017;**2**:239–41.

18. Bhinder B, Gilvary C, Madhukar NS *et al.* Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov* 2021;**11**:900–15.

19. Murphy KP. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

20. Peck RW. Precision Medicine Is Not Just Genomics: The Right Dose for Every Patient. *Annu Rev Pharmacol Toxicol* 2018;**58**:105–22.

21. Tyson RJ, Park CC, Powell JR *et al.* Precision Dosing Priority Criteria: Drug, Disease, and Patient Population Variables. *Frontiers in Pharmacology* 2020;**11**.

22. Polasek TM, Shakib S, Rostami-Hodjegan A. Precision dosing in clinical medicine: present and future. *Expert Review of Clinical Pharmacology* 2018;**11**:743–6.

23. Groenland SL, Verheijen RB, Joerger M *et al.* Precision Dosing of Targeted Therapies Is Ready for Prime Time. *Clinical Cancer Research* 2021;**27**:6644–52.

24. Maxfield K, Milligan L, Wang L *et al.* Proceedings of a Workshop: Precision Dosing: Defining the Need and Approaches to Deliver Individualized Drug Dosing in the Real-World Setting. *Clinical Pharmacology & Therapeutics* 2021;**109**:25–8.

25. Maxfield, Kimberly, Zineh, Issam. Precision Dosing. *JAMA* 2021;**325**:1505–6.

26. Polasek TM, Peck RW. Beyond Population-Level Targets for Drug Concentrations: Precision Dosing Needs Individual-Level Targets that

Include Superior Biomarkers of Drug Responses. *Clinical Pharmacology & Therapeutics* 2024;**n/a**, DOI: 10.1002/cpt.3197.

27. Darwich AS, Polasek TM, Aronson JK *et al.* Model-Informed Precision Dosing: Background, Requirements, Validation, Implementation, and Forward Trajectory of Individualizing Drug Therapy. *Annu Rev Pharmacol Toxicol* 2021;**61**:225–45.

28. 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. *Clinical Pharmacology & Therapeutics* 2017;**101**:646–56.

29. Bartolucci R, Grandoni S, Melillo N *et al. Artificial Intelligence and Machine Learning: Just a Hype or a New Opportunity for Pharmacometrics?*, 2019.

30. Ribba B, Dudal S, Lavé T *et al.* Model-Informed Artificial Intelligence: Reinforcement Learning for Precision Dosing. *Clinical Pharmacology & Therapeutics* 2020;**107**:853–7.

31. Gonzalez D, Rao GG, Bailey SC *et al.* Precision Dosing: Public Health Need, Proposed Framework, and Anticipated Impact. *Clinical and Translational Science* 2017;**10**:443–54.

32. Peck RW. Precision Dosing: An Industry Perspective. *Clin Pharmacol Ther* 2021;**109**:47–50.

33. Neely M, Onufrak N, Scheetz MH *et al.* Supporting Precision Dosing in Drug Labeling. *Clinical Pharmacology & Therapeutics* 2021;**109**:37–41.

34. Wicha SG, Märtson A-G, Nielsen EI *et al.* From Therapeutic Drug Monitoring to Model-Informed Precision Dosing for Antibiotics. *Clinical Pharmacology & Therapeutics* 2021;**109**:928–41.

35. Hawcutt DB, Cooney L, Oni L *et al.* Precision Dosing in Children. *Expert Review of Precision Medicine and Drug Development* 2016;**1**:69–78.

36. Hilmer SN, McLachlan AJ, Le Couteur DG. Clinical pharmacology in the geriatric patient. *Fundamental & Clinical Pharmacology* 2007;**21**:217–30.

37. Buclin T, Thoma Y, Widmer N *et al.* The Steps to Therapeutic Drug Monitoring: A Structured Approach Illustrated With Imatinib. *Frontiers in Pharmacology* 2020;**11**.

38. Chakraborty B, Murphy SA. Dynamic Treatment Regimes. *Annual Review of Statistics and Its Application* 2014;**1**:447–64.

39. Tosca EM, De Carlo A, Ronchi D *et al.* Model-Informed Reinforcement Learning for Enabling Precision Dosing Via Adaptive Dosing. *Clinical Pharmacology & Therapeutics* 2024;**116**, DOI: 10.1002/cpt.3356.

40. Mueller-Schoell A, Groenland SL, Scherf-Clavel O *et al.* Therapeutic drug monitoring of oral targeted antineoplastic drugs. *European Journal of Clinical Pharmacology* 2021;**77**:441–64.

41. Groenland SL, van Eerden RAG, Westerdijk K *et al.* Therapeutic drug monitoring-based precision dosing of oral targeted therapies in oncology: a prospective multicenter study☆. *Annals of Oncology* 2022;**33**:1071–82.

42. Taddeo A, Prim D, Bojescu E-D *et al.* Point-of-Care Therapeutic Drug Monitoring for Precision Dosing of Immunosuppressive Drugs. *The Journal of Applied Laboratory Medicine* 2020;**5**:738–61.

43. Cremers S, Guha N, Shine B. Therapeutic drug monitoring in the era of precision medicine: opportunities! *Br J Clin Pharmacol* 2016;**82**:900–2.

44. Panteli D, Legido-Quigley H, Reichebner C *et al.* Clinical Practice Guidelines as a quality strategy. *Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies [Internet]*. European Observatory on Health Systems and Policies, 2019.

45. Tosca EM, De Carlo A, Bartolucci R *et al.* In silico trial for the assessment of givinostat dose adjustment rules based on the management of key hematological parameters in polycythemia vera patients. *CPT: Pharmacometrics & Systems Pharmacology* 2024;**13**:359–73.

46. Wojciechowski J, Upton RN, Mould DR *et al.* Infliximab Maintenance Dosing in Inflammatory Bowel Disease: an Example for In Silico Assessment of Adaptive Dosing Strategies. *The AAPS Journal* 2017;**19**:1136–47.

47. Sheiner LB. Computer-aided long-term anticoagulation therapy. *Computers and Biomedical Research* 1969;**2**:507–18.

48. Jelliffe RW. Administration of digoxin. *Dis Chest* 1969;**56**:56–60.

49. Keizer RJ, ter Heine R, Frymoyer A *et al.* Model-Informed Precision Dosing at the Bedside: Scientific Challenges and Opportunities. *CPT: Pharmacometrics & Systems Pharmacology* 2018;**7**:785–7.

50. Kantasiripitak W, Van Daele R, Gijsen M *et al.* Software Tools for Model-Informed Precision Dosing: How Well Do They Satisfy the Needs? *Front Pharmacol* 2020;**11**:620.

51. Mizuno T, Dong M, Taylor ZL *et al.* Clinical implementation of pharmacogenetics and model-informed precision dosing to improve patient care. *Br J Clin Pharmacol* 2022;**88**:1418–26.

52. Frymoyer A, Schwenk HT, Zorn Y *et al.* Model-Informed Precision Dosing of Vancomycin in Hospitalized Children: Implementation and Adoption at an Academic Children's Hospital. *Frontiers in Pharmacology* 2020;**11**.

53. Vinks AA, Peck RW, Neely M *et al.* Development and Implementation of Electronic Health Record–Integrated Model-Informed Clinical Decision Support Tools for the Precision Dosing of Drugs. *Clinical Pharmacology & Therapeutics* 2020;**107**:129–35.

54. Klopp-Schulze L, Mueller-Schoell A, Neven P *et al.* Integrated Data Analysis of Six Clinical Studies Points Toward Model-Informed Precision Dosing of Tamoxifen. *Front Pharmacol* 2020;**11**:283.

55. US Food and Drug Administration. Population Pharmacokinetics Guidance for Industry [Online] https://www.fda.gov/media/128793/download. 2022.

56. Upton RN, Mould DR. Basic concepts in population modeling, simulation, and model-based drug development: part 3-introduction to pharmacodynamic modeling methods. *CPT Pharmacometrics Syst Pharmacol* 2014;**3**:e88.

57. Bauer RJ. NONMEM Tutorial Part II: Estimation Methods and Advanced Examples. *CPT: Pharmacometrics & Systems Pharmacology* 2019;**8**:538–56.

58. Karlsson MO, Sheiner LB. The importance of modeling interoccasion variability in population pharmacokinetic analyses. *J Pharmacokinet Biopharm* 1993;**21**:735–50.

59. Jaber MM, Brundage RC. Investigating the contribution of residual unexplained variability components on bias and imprecision of parameter estimates in population pharmacokinetic mixed-effects modeling. *J Pharmacokinet Pharmacodyn* 2023;**50**:123–32.

60. Kaul R, Ossai C, Forkan ARM *et al.* The role of AI for developing digital twins in healthcare: The case of cancer care. *WIREs Data Mining and Knowledge Discovery* 2023;**13**:e1480.

61. T. Erol, A. F. Mendi, D. Doğan. The Digital Twin Revolution in Healthcare. *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. 2020, 1–7.

62. Taylor ZL, Poweleit EA, Paice K *et al.* Tutorial on model selection and validation of model input into precision dosing software for model-informed precision dosing. *CPT: Pharmacometrics & Systems Pharmacology* 2023;**12**:1827–45.

63. Maier C, de Wiljes J, Hartung N *et al.* A continued learning approach for model-informed precision dosing: Updating models in clinical practice. *CPT: Pharmacometrics & Systems Pharmacology* 2022;**11**:185–98.

64. Kluwe F, Michelet R, Mueller-Schoell A *et al.* Perspectives on Model-Informed Precision Dosing in the Digital Health Era: Challenges, Opportunities, and Recommendations. *Clinical Pharmacology & Therapeutics* 2021;**109**:29–36.

65. 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**:607–17.

66. Weinelt FA, Stegemann MS, Theloe A *et al.* Development of a Model-Informed Dosing Tool to Optimise Initial Antibiotic Dosing—A Translational Example for Intensive Care Units. *Pharmaceutics* 2021;**13**, DOI: 10.3390/pharmaceutics13122128.

67. McComb M, Bies R, Ramanathan M. Machine learning in pharmacometrics: Opportunities and challenges. *British Journal of Clinical Pharmacology* 2022;**88**:1482–99.

68. Hutchinson L, Steiert B, Soubret A *et al.* Models and Machines: How Deep Learning Will Take Clinical Pharmacology to the Next Level. *CPT: Pharmacometrics & Systems Pharmacology* 2019;**8**:131–4.

69. Chaturvedula A, Calad-Thomson S, Liu C *et al.* Artificial Intelligence and Pharmacometrics: Time to Embrace, Capitalize, and Advance? *CPT: Pharmacometrics & Systems Pharmacology* 2019;**8**:440–3.

70. Poweleit EA, Vinks AA, Mizuno T. Artificial Intelligence and Machine Learning Approaches to Facilitate Therapeutic Drug Management and Model-Informed Precision Dosing. *Therapeutic Drug Monitoring* 2023;**45**.

71. Coronato A, Naeem M, De Pietro G *et al.* Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine* 2020;**109**:101964.

72. Ribba B. Reinforcement learning as an innovative model-based approach: Examples from precision dosing, digital health and computational psychiatry. *Frontiers in Pharmacology* 2023;**13**.

73. Ribba B, Bräm DS, Baverel PG *et al.* Model enhanced reinforcement learning to enable precision dosing: A theoretical case study with dosing of propofol. *CPT: Pharmacometrics & Systems Pharmacology* 2022;**11**:1497–510.

74. Eckardt J-N, Wendt K, Bornhäuser M *et al.* Reinforcement Learning for Precision Oncology. *Cancers* 2021;**13**, DOI: 10.3390/cancers13184624.

75. Yang C-Y, Shiranthika C, Wang C-Y *et al.* Reinforcement learning strategies in cancer chemotherapy treatments: A review. *Comput Methods Programs Biomed* 2023;**229**:107280.

76. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. MIT press, 2018.

77. Komorowski M, Celi LA, Badawi O *et al.* The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 2018;**24**:1716–20.

78. Liu S, See KC, Ngiam KY *et al.* Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. *J Med Internet Res* 2020;**22**:e18477.

79. De Carlo A, Tosca EM, Fantozzi M *et al.* Reinforcement Learning and PK-PD Models Integration to Personalize the Adaptive Dosing Protocol of Erdafitinib in Patients with Metastatic Urothelial Carcinoma. *Clinical Pharmacology & Therapeutics* 2024;**115**, DOI: 10.1002/cpt.3176.

80. Maier C, Hartung N, Kloft C *et al.* Reinforcement learning and Bayesian data assimilation for model-informed precision dosing in oncology. *CPT: Pharmacometrics & Systems Pharmacology* 2021;**10**:241–54.

81. Yun WJ, Shin M, Jung S *et al.* Deep reinforcement learning-based propofol infusion control for anesthesia: A feasibility study with a 3000-subject dataset. *Comput Biol Med* 2023;**156**:106739.

82. Schamberg G, Badgeley M, Meschede-Krasa B *et al.* Continuous action deep reinforcement learning for propofol dosing during general anesthesia. *Artif Intell Med* 2022;**123**:102227.

83. Moore BL, Pyeatt LD, Kulkarni V *et al.* Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. *The journal of machine learning research* 2014;**15**:655–96.

84. Padmanabhan R, Meskin N, Haddad WM. Optimal adaptive control of drug dosing using integral reinforcement learning. *Math Biosci* 2019;**309**:131–42.

85. Yauney G, Shah P. Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection. In: Doshi-Velez F, Fackler J, Jung K, et al. (eds.). *Proceedings of the 3rd Machine Learning for Healthcare Conference*. Vol 85. PMLR, 2018, 161–226.

86. Yazdjerdi P, Meskin N, Al-Naemi M *et al.* Reinforcement learning-based control of tumor growth under anti-angiogenic therapy. *Computer Methods and Programs in Biomedicine* 2019;**173**:15–26.

87. Ebrahimi Zade A, Shahabi Haghighi S, Soltani M. Reinforcement learning for optimal scheduling of Glioblastoma treatment with Temozolomide. *Computer Methods and Programs in Biomedicine* 2020;**193**:105443.

88. Padmanabhan R, Meskin N, Haddad WM. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Math Biosci* 2017;**293**:11–20.

89. Anzabi Zadeh S, Street WN, Thomas BW. Optimizing warfarin dosing using deep reinforcement learning. *J Biomed Inform* 2023;**137**:104267.

90. Augustin D, Lambert B, Robinson M *et al.* Simulating clinical trials for model-informed precision dosing: using warfarin treatment as a use case. *Frontiers in Pharmacology* 2023;**14**.

91. Tejedor M, Woldaregay AZ, Godtliebsen F. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artif Intell Med* 2020;**104**:101836.

92. Escandell-Montero P, Chermisi M, Martínez-Martínez JM *et al.* Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artif Intell Med* 2014;**62**:47–60.

93. Dean T, Kaelbling LP, Kirman J *et al.* Planning under time constraints in stochastic domains. *Artificial Intelligence* 1995;**76**:35–74.

94. Magni P, Bellazzi R. DT-Planner: an environment for managing dynamic decision problems. *Comput Methods Programs Biomed* 1997;**54**:183–200.

95. Bertsekas D, Tsitsiklis JN. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

96. Zai A, Brown B. *Deep Reinforcement Learning in Action*. Manning Publications, 2020.

97. Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning. Vol 30. 2016.

98. De Carlo, Alessandro, Tosca, Elena Maria, Magni, Paolo. Integrating Reinforcement Learning and PK-PD modelling to enable precision dosing: a multi-objective optimization for the treatment of Polycithemia Vera patients with Givinostat. *PAGE* **31**.

99. Dosne A-G, Valade E, Goeyvaerts N *et al.* Exposure–response analyses of erdafitinib in patients with locally advanced or metastatic urothelial carcinoma. *Cancer Chemotherapy and Pharmacology* 2022;**89**:151–64.

100. Dosne A-G, Valade E, Stuyckens K *et al.* Erdafitinib's effect on serum phosphate justifies its pharmacodynamically guided dosing in patients with cancer. *CPT: Pharmacometrics & Systems Pharmacology* 2022;**11**:569–80.

101. Dosne A-G, Valade E, Stuyckens K *et al.* Population Pharmacokinetics of Total and Free Erdafitinib in Adult Healthy Volunteers and Cancer Patients: Analysis of Phase 1 and Phase 2 Studies. *The Journal of Clinical Pharmacology* 2020;**60**:515–27.

102. Tabernero J, Bahleda R, Dienstmann R *et al.* Phase I Dose-Escalation Study of JNJ-42756493, an Oral Pan–Fibroblast Growth Factor Receptor Inhibitor, in Patients With Advanced Solid Tumors. *JCO* 2015;**33**:3401–8.

103. US Food and Drug Administration. FDA approves first targeted therapy for metastatic bladder cancer. *FDA* 2020.

104. FDA. Erdafitnib Prescribing Information.

105. Bahleda R, Italiano A, Hierro C *et al.* Multicenter Phase I Study of Erdafitinib (JNJ-42756493), Oral Pan-Fibroblast Growth Factor Receptor Inhibitor, in Patients with Advanced or Refractory Solid Tumors. *Clinical Cancer Research* 2019;**25**:4888–97.

106. Yanochko GM, Vitsky A, Heyen JR *et al.* Pan-FGFR Inhibition Leads to Blockade of FGF23 Signaling, Soft Tissue Mineralization, and Cardiovascular Dysfunction. *Toxicological Sciences* 2013;**135**:451–64.

107. Loriot Y, Necchi A, Park SH *et al.* Erdafitinib in Locally Advanced or Metastatic Urothelial Carcinoma. *N Engl J Med* 2019;**381**:338–48.

108. Lawrence Gould A, Boye ME, Crowther MJ *et al.* Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in medicine* 2015;**34**:2181–95.

109. Monahan GE. State of the Art—A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms. *Management Science* 1982;**28**:1–16.

110. Chifotides HT, Bose P, Verstovsek S. Givinostat: an emerging treatment for polycythemia vera. *Expert Opinion on Investigational Drugs* 2020;**29**:525–36.

111. Iurlo A, Cattaneo D, Bucelli C *et al.* New Perspectives on Polycythemia Vera: From Diagnosis to Therapy. *International Journal of Molecular Sciences* 2020;**21**, DOI: 10.3390/ijms21165805.

112. Marchioli R, Finazzi G, Landolfi R *et al.* Vascular and Neoplastic Risk in a Large Cohort of Patients With Polycythemia Vera. *JCO* 2005;**23**:2224–32.

113. Stein BL, Patel K, Scherber R *et al.* MPN-133: Mortality and Causes of Death of Patients with Polycythemia Vera: Analysis of the REVEAL Prospective, Observational Study. *Clinical Lymphoma Myeloma and Leukemia* 2021;**21**:S355.

114. Hochhaus A, Baccarani M, Silver RT *et al.* European LeukemiaNet 2020 recommendations for treating chronic myeloid leukemia. *Leukemia* 2020;**34**:966–84.

115. MATLAB. *R2021a*. Natick, Massachussetts. The MathWorks Inc., 2021.

116. Schmidinger M, Danesi R, Jones R *et al.* Individualized dosing with axitinib: rationale and practical guidance. *Future Oncol* 2018;**14**:861–75.

117. Larkin J, Fishman M, Wood L *et al.* Axitinib for the treatment of metastatic renal cell carcinoma: recommendations for therapy management to optimize outcomes. *Am J Clin Oncol* 2014;**37**:397–403.

118. Parkin DM, Bray F, Ferlay J *et al.* Global Cancer Statistics, 2002. *CA: A Cancer Journal for Clinicians* 2005;**55**:74–108.

119. Motzer RJ, Bander NH, Nanus DM. Renal-cell carcinoma. *N Engl J Med* 1996;**335**:865–75.

120. Ljungberg B, Bensalah K, Canfield S *et al.* EAU guidelines on renal cell carcinoma: 2014 update. *Eur Urol* 2015;**67**:913–24.

121. Rothermundt C, von Rappard J, Eisen T *et al.* Second-line treatment for metastatic clear cell renal cell cancer: experts' consensus algorithms. *World J Urol* 2017;**35**:641–8.

122. Pfizer Laboratories Div Pfizer Inc. INLYTA — axitinib tablet, film coated: Full prescribing information.

123. Bhargava P. VEGF kinase inhibitors: how do they cause hypertension? *Am J Physiol Regul Integr Comp Physiol* 2009;**297**:R1-5.

124. Jain M, Townsend RR. Chemotherapy agents and hypertension: a focus on angiogenesis blockade. *Curr Hypertens Rep* 2007;**9**:320–8.

125. Rini BI, Melichar B, Fishman MN *et al.* Axitinib dose titration: analyses of exposure, blood pressure and clinical response from a randomized phase II study in metastatic renal cell carcinoma. *Ann Oncol* 2015;**26**:1372–7.

126. Rini BI, Garrett M, Poland B *et al.* Axitinib in metastatic renal cell carcinoma: results of a pharmacokinetic and pharmacodynamic analysis. *J Clin Pharmacol* 2013;**53**:491–504.

127. Eto M, Uemura H, Tomita Y *et al.* Overall survival and final efficacy and safety results from a Japanese phase II study of axitinib in cytokine-refractory metastatic renal cell carcinoma. *Cancer Sci* 2014;**105**:1576–83.

128. Tomita Y, Uemura H, Fujimoto H *et al.* Key predictive factors of axitinib (AG-013736)-induced proteinuria and efficacy: a phase II study in Japanese patients with cytokine-refractory metastatic renal cell Carcinoma. *Eur J Cancer* 2011;**47**:2592–602.

129. Chen Y, Rini BI, Bair AH *et al.* Population pharmacokinetic-pharmacodynamic modelling of 24-h diastolic ambulatory blood pressure changes mediated by axitinib in patients with metastatic renal cell carcinoma. *Clin Pharmacokinet* 2015;**54**:397–407.

130. Rini BI, Quinn DI, Baum M *et al.* Hypertension among patients with renal cell carcinoma receiving axitinib or sorafenib: analysis from the randomized phase III AXIS trial. *Target Oncol* 2015;**10**:45–53.

131. Keizer RJ, Gupta A, Mac Gillavry MR *et al.* A model of hypertension and proteinuria in cancer patients treated with the anti-angiogenic drug E7080. *J Pharmacokinet Pharmacodyn* 2010;**37**:347–63.

132. Wilhelm MP. Vancomycin. *Mayo Clin Proc* 1991;**66**:1165–70.

133. Ye Z-K, Tang H-L, Zhai S-D. Benefits of therapeutic drug monitoring of vancomycin: a systematic review and meta-analysis. *PloS one* 2013;**8**:e77169.

134. Vancomycin. *Drugs and Lactation Database (LactMed®)*. Bethesda (MD): National Institute of Child Health and Human Development, 2006.

135. Garreau R, Falquet B, Mioux L *et al.* Population Pharmacokinetics and Dosing Simulation of Vancomycin Administered by Continuous Injection in Critically Ill Patient. *Antibiotics (Basel)* 2021;**10**, DOI: 10.3390/antibiotics10101228.

136. Cardiff Critical Care Department. Intensive Care Units - Guidelines for Vancomycin by continuous infusion.

137. Roberts JA, Taccone FS, Udy AA *et al.* Vancomycin dosing in critically ill patients: robust methods for improved continuous-infusion regimens. *Antimicrob Agents Chemother* 2011;**55**:2704–9.

138. Thijs L, Quintens C, Vander Elst L *et al.* Clinical Efficacy and Safety of Vancomycin Continuous Infusion in Patients Treated at Home in an

Outpatient Parenteral Antimicrobial Therapy Program. *Antibiotics (Basel)* 2022;**11**, DOI: 10.3390/antibiotics11050702.

139. Hughes JH, Long-Boyle J, Keizer RJ. Maximum a posteriori Bayesian methods out-perform non-compartmental analysis for busulfan precision dosing. *Journal of Pharmacokinetics and Pharmacodynamics* 2024, DOI: 10.1007/s10928-024-09915-w.

140. Le Louedec F, Puisset F, Thomas F *et al.* Easy and reliable maximum a posteriori Bayesian estimation of pharmacokinetic parameters with the open-source R package mapbayr. *CPT Pharmacometrics Syst Pharmacol* 2021;**10**:1208–20.

141. Leven C, Coste A, Mané C. Free and Open-Source Posologyr Software for Bayesian Dose Individualization: An Extensive Validation on Simulated Data. *Pharmaceutics* 2022;**14**, DOI: 10.3390/pharmaceutics14020442.

142. Maier C, Hartung N, de Wiljes J *et al.* Bayesian Data Assimilation to Support Informed Decision Making in Individualized Chemotherapy. *CPT Pharmacometrics Syst Pharmacol* 2020;**9**:153–64.

143. Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 1992;**8**:229–56.

144. Coulom R. Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. In: van den Herik HJ, Ciancarini P, Donkers HHLM (Jeroen) (eds.). *Computers and Games*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, 72–83.

145. Kocsis L, Szepesvari C. Bandit Based Monte-Carlo Planning. *European Conference on Machine Learning*. 2006.

146. Silver D. Reinforcement learning and simulation-based search in computer go. 2009.

147. Mnih V, Kavukcuoglu K, Silver D *et al.* Human-level control through deep reinforcement learning. *Nature* 2015;**518**:529–33.

148. Anenberg B, Raghavan B. Sampling Strategies for Deep Reinforcement Learning. *arXiv preprint arXiv:13125602* 2013.

149. Yang Z, Xie Y, Wang Z. A Theoretical Analysis of Deep Q-Learning. *Conference on Learning for Dynamics & Control*. 2019.

150. Wang Z, Schaul T, Hessel M *et al.* Dueling network architectures for deep reinforcement learning. PMLR, 2016, 1995–2003.

151. Zhang S, Wu Y, Ogai H *et al.* Tactical decision-making for autonomous driving using dueling double deep Q network with double attention. *IEEE Access* 2021;**9**:151983–92.

152. Mehta D. State-of-the-art reinforcement learning algorithms. *International Journal of Engineering Research and Technology* 2020;**8**:717–22.

153. Kwon Y, Saltaformaggio B, Kim IL *et al.* A2c: Self destructing exploit executions via input perturbation. 2017.

154. Lillicrap TP, Hunt JJ, Pritzel A *et al.* Continuous control with deep reinforcement learning. *arXiv preprint arXiv:150902971* 2015.

155. Conti E, Madhavan V, Petroski Such F *et al.* Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. *Advances in neural information processing systems* 2018;**31**.

156. Friberg LE, Henningsson A, Maas H *et al.* Model of chemotherapy-induced myelosuppression with parameter consistency across drugs. *J Clin Oncol* 2002;**20**:4713–21.

157. Schindler E, Amantea M, Karlsson M *et al.* A Pharmacometric Framework for Axitinib Exposure, Efficacy, and Safety in Metastatic Renal Cell Carcinoma Patients. *CPT: Pharmacometrics & Systems Pharmacology* 2017;**6**:373–82.

158. WHO. WHOCC - ATC/DDD Index.