Clinical Pharmacokinetics (2024) 63:1221–1237
https://doi.org/10.1007/s40262-024-01409-9

REVIEW ARTICLE

# Machine Learning Methods for Precision Dosing in Anticancer Drug Therapy: A Scoping Review

Olga Teplytska[1] · Moritz Ernst[2] · Luca Marie Koltermann[1] · Diego Valderrama[3] · Elena Trunz[4] · Marc Vaisband[5,6,7] · Jan Hasenauer[5,6] · Holger Fröhlich[3,8] · Ulrich Jaehde[1]

Accepted: 4 August 2024 / Published online: 17 August 2024
© The Author(s) 2024

## Abstract

**Introduction** In the last decade, various Machine Learning techniques have been proposed aiming to individualise the dose of anticancer drugs mostly based on a presumed drug effect or measured effect biomarkers. The aim of this scoping review was to comprehensively summarise the research status on the use of Machine Learning for precision dosing in anticancer drug therapy.

**Methods** This scoping review was conducted in accordance with the interim guidance by Cochrane and the Joanna Briggs Institute. We systematically searched the databases Medline (via PubMed), Embase and the Cochrane Library for research articles and reviews including results published after 2016. Results were reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist.

**Results** A total of 17 relevant studies was identified. In 12 of the included studies, Reinforcement Learning methods were used, including Classical, Deep, Double Deep and Conservative Q-Learning and Fuzzy Reinforcement Learning. Furthermore, classical Machine Learning methods were compared in terms of their performance and an artificial intelligence platform based on parabolic equations was used to guide dosing prospectively and retrospectively, albeit only in a limited number of patients. Due to the significantly different algorithm structures, a meaningful comparison between the various Machine Learning approaches was not possible.

**Conclusion** Overall, this review emphasises the clinical relevance of Machine Learning methods for anticancer drug dose optimisation, as many algorithms have shown promising results enabling model-free predictions with the potential to maximise efficacy and minimise toxicity when compared to standard protocols.

**Key Points**

Machine Learning has great potential to advance PK/PD-guided dosing strategies in oncology.

Reinforcement Learning methods are increasingly developed for automated dose individualisation of anticancer drugs.

## 1 Introduction

Recently, various Machine Learning (ML) techniques have been explored that aim to individualise dosing, based on a presumed drug effect, measured effect biomarkers or pharmacokinetic (PK) measurements. Based on 86 studies identified in a pilot literature search from our working group, the most frequently used algorithms for dose optimisation were Decision Trees and their ensembles, such as Random Forests and Boosting Algorithms, Support Vector Machines and Artificial Neural Networks. Specifically, Reinforcement Learning (RL) has played a significant role. Mostly, the identified algorithms aided in the dose individualisation of anticoagulants, immunosuppressants and antibiotics [1–5]. Furthermore, ML was applied to problems in pharmacometrics including PK and pharmacodynamic (PD) modelling and simulation, model-informed precision dosing, and systems pharmacology [6]. In radiotherapy, ML methods were successfully implemented for synthetic computed tomography (sCT) image generation [7–9], auto-segmentation to achieve a more accurate tumour delineation [10–12] and knowledge-based treatment planning, with the aim to deliver an acceptable dose into the target organ while

Extended author information available on the last page of the article

sparing surrounding organs at risk [13–15]. While much has already been published on radiotherapy dose optimisation, there is currently no review comprehensively summarising the research status on the use of ML for precision dosing in anticancer drug therapy. As predicting and improving treatment outcomes through treatment individualisation is a major goal and challenge in oncology, and ML can be an asset to this goal, we conducted a scoping review on this topic to summarise the applications and discuss their benefits and limitations. Most publications we identified were in the field of RL, highlighting the importance of this method for dose optimisation in oncology. Therefore, this review mainly focuses on RL methods.

## 2 Methods

This scoping review was conducted in accordance with the interim guidance by Cochrane and the Joanna Briggs Institute [16]. Results were reported based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist [17]. Studies published from 2016, that is, after the open-source launch of the ML framework TensorFlow, were included. This library, which was developed by the Google-Brain-Team, encouraged many scientists, including life science researchers, who were new to ML, to join this research field. PyTorch, which is another popular ML framework developed by Meta AI (then Facebook Inc.), was initially released in September 2016. We hypothesized that the development of these open-source libraries and their subsequent ubiquitous adoption meant that results published before and after 2016 would probably not be comparable. The databases Medline (via PubMed), Embase and the Cochrane Library were searched for primary trials and evidence syntheses on 3 March 2023.

The search strategy for primary studies and the search strategy for evidence syntheses are shown in Online Resources 1 and 2, respectively. Only studies in English language that focused on dose optimisation of anticancer drugs in human patients and reported quantitative results on model performance, dose optimisation parameters and/or resulting doses were included. The full inclusion and exclusion criteria are shown in Table 1. Two authors (OT and ME) independently screened titles and abstracts of identified records and retrieved full-text articles of potentially eligible studies, using the automated tool Rayyan for screening [18]. All reviews and included publications were thoroughly searched for further eligible studies (citation searching). Subsequently, they independently assessed eligibility of the remaining records using the pre-defined eligibility criteria. In the case of disagreement, a third author (DV) adjudicated. Studies on radiotherapy dose optimisation were excluded at the end of the screening process.

Afterwards, study methods, population specifics and outcomes, as well as conflict of interest and reproducibility parameters (whether data, code and model topology were reported) were extracted by the first author (OT) using a standardised form. Another author (LMK) independently reviewed the data extraction. In addition, a non-systematic keyword search without double-screening was carried out in PubMed, Embase and ResearchGate on 21 February 2024 using the keywords from the original search strategy and the keyword "Reinforcement Learning" because most of the studies identified in the systematic search were located in this field. This was done to avoid missing the most recent publications. The objectives, inclusion criteria and methods for this scoping review were specified in advance and documented in a protocol including documentation of protocol adjustments [19]. As per protocol, if the original code and/or dataset were not reported, the authors were contacted between January and February 2024, and the obtained

**Table 1** Predefined inclusion and exclusion criteria

| | Inclusion criteria | Exclusion criteria |
|---|---|---|
| Study designs | Randomised controlled trials<br>Observational studies (case control, cohort, cross-sectional)<br>Prospective and retrospective design<br>Non-comparative/non-controlled single-arm studies with any number of participants<br>Case studies with any number of participants<br>Rapid/Living/Scoping/SystematicReviews/Meta-analyses<br>Modelling studies | Commentaries, editorials, empirical studies and other publications without quantitative study results |
| Population | Patients treated for any type of cancer | |
| Setting | Studies reported from 01/2016 in English language | |
| Interventions | Machine Learning algorithms aiming at optimising the dose of anticancer drug agents | |
| Outcomes | Quantitative results on model performance and dose optimisation parameters and/or resulting doses | |

information was assessed and added. Finally, the search results were displayed in a flowchart according to PRISMA criteria [20] (Fig. 1) and the results were reported in comprehensive tables which display all relevant study characteristics (Tables 2, 3) and quality criteria (Online Resources 3 and 4).

## 3 Results

A total of 2024 publications was found in databases and a further 7 were identified through citation searching of reviews (Fig. 1). Thirty-one duplicates (8 identified by automatic deduplication) and 7 publications not written in English or published before 2016 were excluded and 1986 publications were screened. After title and abstract screening, 191 studies remained for full text screening, of which 153 could be retrieved. Among these studies, 14 focused on dose optimisation in drug therapy and were included in this review. The other studies were excluded because they focussed on radiotherapeutic dose optimisation ($n = 111$), did not focus on dose optimisation ($n = 21$), did not report results ($n = 9$) or did not apply ML methods ($n = 5$). Three studies, published after the systematic search was conducted, were additionally identified in a non-systematic search on 21 February 2024. Finally, 17 studies were included in this review and the results were categorised according to the methods used (RL methods, classical ML and Phenotypic Personalised Medicine).

An overview of the identified methods can be found in Fig. 2 and short summaries of the most important study features and quality criteria are depicted in Tables 2 and 3 and Online Resources 3 and 4, respectively.

### 3.1 Reinforcement Learning

From the 17 studies that were included in the review, 12 used RL for drug dose optimisation. Reinforcement Learning is a ML technique, in which an agent learns in an interactive environment by trial and error using feedback (reward or punishment) from its own actions and experiences, in a goal-driven manner. The goal in RL is to find a suitable action model, which maximises the total cumulative reward. For example, in the context of oncology, one might typically aim to alter a chemotherapeutic regimen to reach maximal tumour shrinkage. In many of the identified publications, the model parameters were perturbed to assess robustness of the model, which we have summarised in Table 2.

Additional to the studies included in this review, we identified two comprehensive reviews that summarised literature

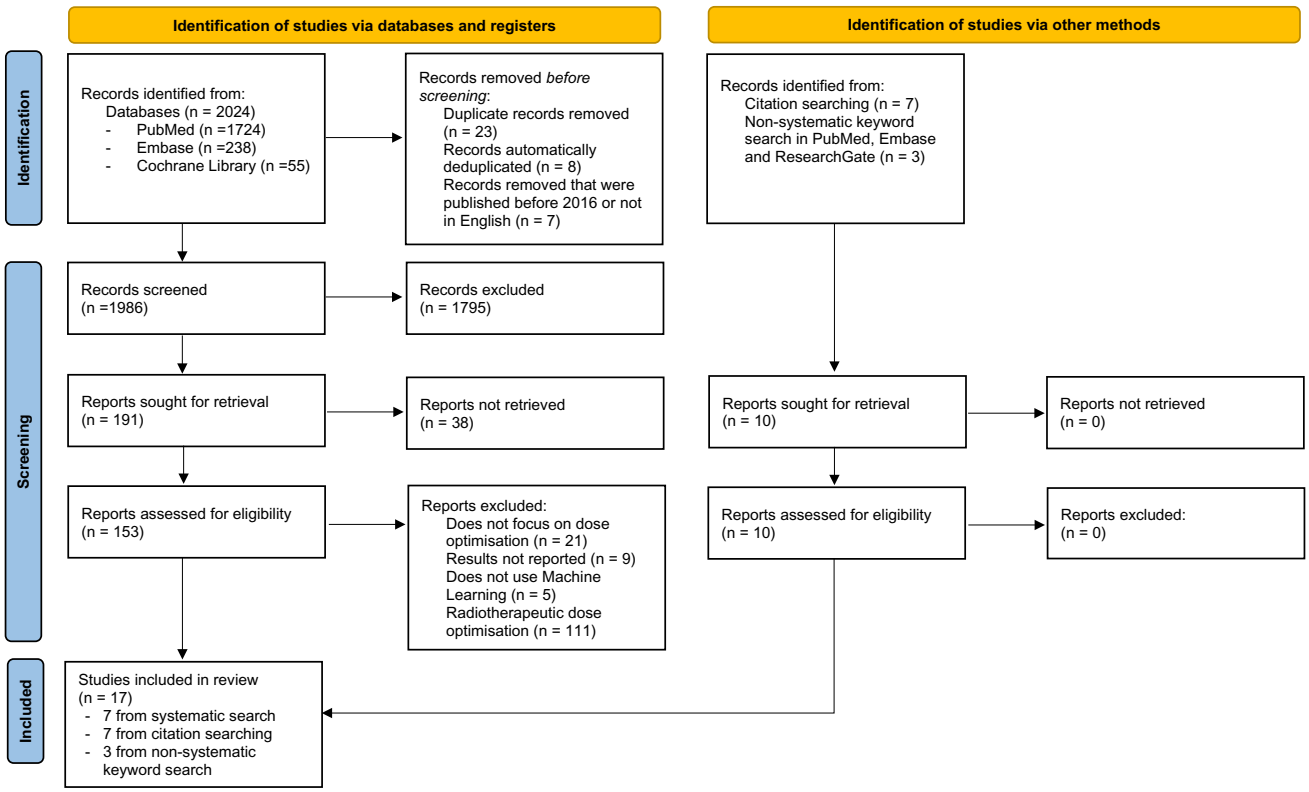



**Fig. 1** Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram including searches of databases, registers and other sources [20]

**Table 2** Summary of methodology and outcomes (Reinforcement Learning)

| Study | Population | Methods | Model parameter perturbation or repetition of runs (robustness) | Outcome and details of comparison |
|---|---|---|---|---|
| **Classical Q-Learning** | | | | |
| Yazdjeri et al. [25] | Simulated patients treated with endostatin iv<br>(1) max. 50 mg/kg/day<br>(2) max. 30 mg/kg/day | Reward: Reduce tumour size to desired volume of $\leq 1$ mm$^3$, while $\leq 2$ mm$^3$ is still considered acceptable | x | Scenario (1)<br>- Designed controller faster than literature models<br>- Could reach smaller final tumour volume (7.774 mm$^3$ vs 69 mm$^3$ vs 0.12 mm$^3$) with halved drug dose<br>Scenario (2)<br>- Total amount of inhibitor, maximal dose and final tumour volume considerably lower<br>- Total amount, maximal dose and tumour volume improved<br>- Amount of injected endostatin nearly equal to scenario 1 but tumour shrinkage took more time |
| Ebrahimi Zade et al [29] | Synthetic glioblastoma patient weighting 70 kg treated with TMZ<br>3 different tumour entities with different cell counts<br>(1) 50 k white<br>(2) 30 k white<br>(3) 70 k grey simulated 10 times | Reward:Minimise tumour size, optimise TMZ schedule considering tumour size Effects, tumour removal and patient death<br>- Tumour growth model feeds into RL optimisation model<br>- Max. 2100 mg in 28 days administered<br>- Compared to classical 7/14 regimen<br>Case (1) 150 mg TMZ or none, for entity (1)<br>Case (2) 150 mg, 200 mg or none, for entities (2, 3) | x | - RL model more successful in optimising the reward than classical 7/14 regimen<br>- Optimal schedule: 150 mg of TMZ every other day<br>- Different schedules depending on tumour entities and cell count: (1, 2) 150 mg every other day (3) starting with 3×200 mg and after break 150 mg every other day |
| De Carlo et al. [31] | 141 patients with metastatic urothelial carcinoma treated with erdafitinib simulated<br>(1) 96 complete responders<br>(2) 45 partial responders | Reward: Maintain serum phosphate concentration within target range of 5.5–7 mg/dL<br>- Possible doses available in practice: 4, 5, 6, 8, 9 mg/day<br>- Method applied to personalise 5-month treatment period, compared to FDA protocol<br>- Assessment after two weeks, at the end of the fourth month and at the end of treatment | – | - Q-Learning based protocols: Higher percentage of patients within target range, lower percentage of hyperphosphatemia<br>- End of the fourth month:<br>Complete responders: Q-Learning-based protocols - 98% successful (vs 70.21% for the FDA protocols)<br>- End of treatment: Partial responders: Q-Learning–based protocols - 68.09% successful (vs 56.73%)<br>- Q-Learning proposed lower starting dose than FDA protocol and gradual increase for both groups |

**Table 2** (continued)

| Study | Population | Methods | Model parameter perturbation or repetition of runs (robustness) | Outcome and details of comparison |
|---|---|---|---|---|
| Padmanabhan et al. [35–37] | Three cases simulated for 15 patients:<br>(1) Adult<br>(2) Pregnant woman<br>(3) Multimorbid elderly person<br>General model for iv applied drug | Four states representing numbers of immune, normal and tumour cells and drug concentrations<br>Rewards:<br>(1) Eradicate tumour cells aggressively<br>(2) Minimise drug concentration. before childbirth, after childbirth as (1)<br>(3) Keep normal cell count high during therapy due to comorbidities | x | Optimisation successful after 28 days of simulation on average |
| **Deep Q-Learning** | | | | |
| Yauney et al. [41] | 50 simulated glioma patients receiving<br>(1) Temozolomide<br>(2) Procarbazine, nitrosourea, CCNU and vincristine | Reward: Reduce mean tumour diameter, in some experiments additionally: Lower dose to avoid toxicity<br>RL Deep Neural Network interacted with tumour growth inhibition model<br>(1) Patient-based experiments<br>(2) Trial-based experiments<br>Fixed concentration trials:<br>- 200 mg/m$^2$/d of TMZ<br>- PCV: 60 mg/m$^2$ of procarbazine on Days 8 to 21, 110 mg/m$^2$ of CCNU on Day 1, and 1.4 mg/m$^2$ of vincristine on Days 8 and 29 of each cycle<br>Variable concentration trials: administration of 25, 50, 75 or 100% of the fixed dose or none | – | - All algorithms able to predict optimal regimens<br>- RL equally good at reducing mean tumour diameter as expert opinions unless only 25% of the fixed dose applied |
| Eastman et al. [45] | 200 simulated patients<br>Model of breast cancer and ovarian cancer growth in mice applied<br>General model, not validated in humans | Reward: Minimise tumour cell count while reducing toxicity using information on relative BMD<br>Method compared to traditional nominal optimal controller and nearest testing neighbour optimal controller after adding information on relative BMD | x | (1) Comparison with traditional nominal optimal controller:<br>- RL agent produced schedules closer to theoretical optimum<br>- Benefit of using RL increased with strength of parameter perturbation<br>(2) Comparison with nearest testing neighbour optimal controller:<br>- RL controller outperformed other method at 20 and 25% perturbation strength level, but not 15% level |

**Table 2** (continued)

| Study | Population | Methods | Model parameter perturbation or repetition of runs (robustness) | Outcome and details of comparison |
|---|---|---|---|---|
| Huo et al. [48] | (1) Patient in good health → immune cell threshold 50% from baseline<br>(2) Patient in poor condition → immune cell threshold 70% from baseline<br>General model | Reward: Maintain number of effector-immune cells and administered drug dose within certain range while minimising tumour cell count<br>MIER-MO-DQN compared to conventional DQN approach and linear weighted sum function-based DQN (W_DQN) | x | MIER-MO-DQN:<br>- More effective personalised plans: Tumour cell counts lower (18.7 for patient in good health and 31.6% for patient in poor condition), more efficient (treatment time reduced to 1/3), lower maintenance doses<br>- More fluctuation in earlier iterations, but converged more quickly later than through random replay |
| Mashayekhi et al. [49] | Young, pregnant and elderly patient simulated with different parameter values:<br>(1) Young adult patient<br>(2) Pregnant woman<br>(3) Elderly patient<br>General model for iv applied drug | Four states representing numbers of immune, normal and tumour cells and drug concentrations<br>Rewards:<br>(1) Eradicate tumour cells aggressively<br>(2) Minimise drug concentration<br>Method compared to simple Q-Learning method from Padmanabhan et al and non-RL controllers | x | Method showed similar trend as simple Q-Learning (Padmanabhan et al.) from different aspects but with reductions in<br>- Treatment period<br>- Exposure: AUC reduced by approximately 1–3 mg·d/L, in one case 0.27 mg·d/L higher<br>- Time to reach goal state (by 1–4 days)<br>- Compared to non-RL controllers, the total drug dose is smaller except for when compared to a state feedback control strategy |
| **Fuzzy reinforcement learning** | | | | |
| Treesatayapun et al. [50, 51] | Four patients with different responsiveness to drug therapy<br>General model: No exact drug or tumour entity | Reward: Complete eradication of tumour cells | – | - More sensitive patients received lower doses<br>- Normal and immune cells dropped in the beginning of therapy but recovered after approximately 40 days for all cases |
| Alsaadi et al. [56] | Young and elderly patient simulated with different parameter values:<br>(1) Young patient, no uncertainty modelling<br>(2) Elderly patient, no uncertainty modelling<br>(3) Proposed controller and Watkin's Q-Learning method exerted to ten different patients and compared, with uncertainty modelling<br>General model | Reward:<br>Complete eradication of tumour cells for young patients, control normal and tumour cells counts for elderly patients (avoid toxicity) | x | (1) Tumour cells reduced to 0 with acceptable initial drop in normal and immune cells and subsequent recovery<br>(2) Tumour cells reduced, but not to 0. Initial drop in normal and immune cells, but faster recovery than in (1). 5.5% lower amount of drug administered than in (1)<br>Superiority of proposed method compared to Q-Learning:<br>- Faster: 500 vs 50000 epochs<br>- Error reduced by 35 and 24%<br>- Drug amount reduced by 1 and 10% |

**Table 2** (continued)

| Study | Population | Methods | Model parameter perturbation or repetition of runs (robustness) | Outcome and details of comparison |
|---|---|---|---|---|
| **Other methods** | | | | |
| Maier et al. [57] | 1000 simulated patients suffering from NSCLC small-cell lung carcinoma<br>Single dose of paclitaxel-based chemotherapy every 3 weeks for 6 cycles | Bayesian DA-RL<br>Reward: Optimise neutropenia grade to avoid grades 0 and 4 to achieve optimal efficacy and minimal toxicity<br>Comparison of:<br>(1) Standard dosing (200 mg/m$^2$ BSA and 20% dose reduction for grade 4 neutropenia)<br>(2) PK-guided dosing<br>(3) MAP-guided dosing<br>(4) Bayesian DA with individualised uncertainty quantification<br>(5) Bayesian data assimilation-Reinforcement Learning | x | - PK-guided: Nadir concentrations not as low as with standard dosing but increased incidence of grade 0 neutropenia<br>- MAP-guided dosing: Incidence of grade 4 neutropenia accumulated<br>- DA consistently maintained nadir concentrations within target range with reduced variability<br>- Incidence of grade 0 and 4 neutropenia significantly reduced in later cycles<br>- RL-guided dosing able to control the neutrophil concentration well throughout cycles<br>- DA-RL: Combining individualised uncertainties and patient states kept nadir concentrations in the target range with reduced variability |
| Shiranthika et al. [59] | 40 stage 4 colon carcinoma patients receiving first-line chemotherapy, bevacizumab-based or cetuximab-based | Conservative Q-Learning using a SOCR<br>Reward: Includes changes in tumour size, body weight, drug response, overall side effects, and patient death<br>Proposed schedule supervised by considering previous treatment decisions of oncologists | – | - Suggested therapy consistent in-between cycles<br>- RMSE differences in an acceptable range (8.14–83.53) and comparatively small when Relative Doses Index was similar |

×: method applied, –: method not applied

*AUC* area under the drug plasma concentration-time curve, *BMD* bone marrow density, *BSA* body surface area, *CCNU 1*-(2-chloroethyl)-3-cyclohexyl-1-nitrosourea, or: lomustine, *DA* data assimilation, *DQN* Deep Q-Network, *FDA* U.S. Food and Drug Administration, *iv* intravenous, *MAP* maximum a posteriori, *MIER-MO-DQN* Multi-Objective Deep Q-Network based on Multi Indicator Experience Replay, *NSCLC* non–small-cell lung carcinoma, *PK* pharmacokinetic, *RMSE* root mean squared error, *RL* Reinforcement Learning, *SOCR* supervised optimal chemotherapy regimen, *TMZ* temozolomide, *W_DQN* linear weighted sum function-based DQN

**Table 3** Summary of methodology and outcomes (other than Reinforcement Learning)

| Study | Population | Methods | Cross validation | Outcome and details of comparison |
|---|---|---|---|---|
| **Classical Machine Learning approaches** | | | | |
| Kozłowska et al. [60] | 42 NSCLC patients: 17 untreated, 25 treated with platinum-based doublet chemotherapy<br>Simulated therapy regimens:<br>(A) MTD<br>(B) MTD with drug holidays<br>(C) MT: Dose smaller than in MTD but more frequent administration<br>(D) MT with drug holidays | Computational platform combining Machine Learning and mathematical model<br>Goal: Predict response to chemotherapy and find best regimen for palliative patients<br>- Mathematical model calibrated with a multivariate Gaussian mixture model using an expectation-maximisation method<br>- Competition between cancer cells sensitive and resistant to chemotherapy modelled | – | If competition between sensitive and resistant cells considered low: No difference between schedules with or without drug holidays<br>If competition assumed to be high: Best<br>outcome in schedules including drug holidays (amount of drug sensitive cells increases during that time) |
| Yu et al. [62] | 149 HER2(+) breast cancer patients<br>(1) 55 patients (HR(−), metastatic) lapatinib dose of 1000 mg<br>(2) 94 patients (advanced disease) lapatinib dose of 1250 mg | Sequential forward selection algorithm based on Random Forests to select minimal size of feature subset<br>Logistic Regression, Support Vector Machine, Random Forest, Adaboost, XGBoost, Gradient Boosted Decision Trees, LightGBM, CatBoost, TabNet, Neural Network, Super TML and Wide & Deep algorithms for dose prediction | x | Most important features: Weight, number of previous chemotherapy treatments and metastases and especially underlying treatment protocol<br>Most accurate algorithm for dose prediction: TabNet (accuracy of 0.82 and AUC of 0.83):<br>(1) precision of 88% and recall rate of 64%<br>(2) precision of 82% and recall rate of 95% |
| Cauvin et al. [63] | 80 adult patients with head and neck with cisplatin-based chemotherapy: 3-h iv infusion of 75–100 mg/m$^2$ cisplatin repeated every 21 days for a minimum of three cycles | Naive Bayes, Logistic Regression, Neural Network, Gradient Boosted Trees, Decision Tree, Random Forest, XGBoost Trees and a Generalized Linear Model used to identify the exposure parameter most strongly associated with response | x | Most accurate algorithm: Generalized Linear Model with accuracy of 0.71<br>$C_{max}$: Most useful exposure parameter to explain response (accuracy reduction: 0.062)<br>Proposed $C_{max}$ range: 2.1–4.1 µg/mL<br>PK-guided dosing: Four patients with progressive disease would have received higher doses and three patients with alleviated toxicity would have received lower doses compared to actual dosing |
| **Phenotypic personalised medicine (PPM) case studies** | | | | |
| Lee et al. [67] | Two paediatric patients with standard risk ALL on maintenance therapy<br>(1) Four drug regimens (dexamethasone, vincristine, mercaptopurine, methotrexate)<br>(2) Two drug regimens (mercaptopurine, methotrexate) | CURATE.AI goal:<br>- ANCs between 500 and 1500/µL<br>- Platelet counts above 50,000/µL | – | CURATE.AI: Better ANC ranges for both examined patients and both regimens compared to experimental regimens<br>Dose proposed by CURATE.AI usually significantly lower than experimental dose |
| Pantuck et al. [68] | 82-year-old male patient with mCRPC, chemotherapy of enzalutamide and ZEN 3694 (bromodomain inhibitor)<br>Starting doses: 160 mg enzalutamide and 48 mg ZEN 3694 once daily | CURATE.AI recommended dose combinations of enzalutamide and ZEN 3694 to minimise the patient's PSA level<br>6 months of physician-guided treatment followed by 6 months of CURATE.AI recommendations | – | Final dose: 120 mg of enzalutamide, 24 mg of ZEN-3964<br>PSA level of 0.78 ng/mL (14.3 ng/mL in the beginning; 0.91 ng/mL after physician-guided dosing) |

×: method applied; −: method not applied/not specified

*ALL* acute lymphoblastic leukaemia, *ANCs* absolute neutrophil counts, *AUC* area under the curve, $C_{max}$ maximum concentration, *HER2(+)* human epidermal growth factor receptor 2 positive breast cancer, *HR(−)* hormone receptor negative breast cancer, *iv* intravenous, *LightGBM* Light Gradient-Boosting Machine, *mCRPC* metastatic castration resistant prostate cancer, *MT* metronomic treatment, *MTD* maximal tolerated dose, *NSCLC* non-small-cell lung cancer, *PPM* phenotypic personalised medicine, *PSA* prostate-specific antigen, *XGBoost* Extreme Gradient Boosting

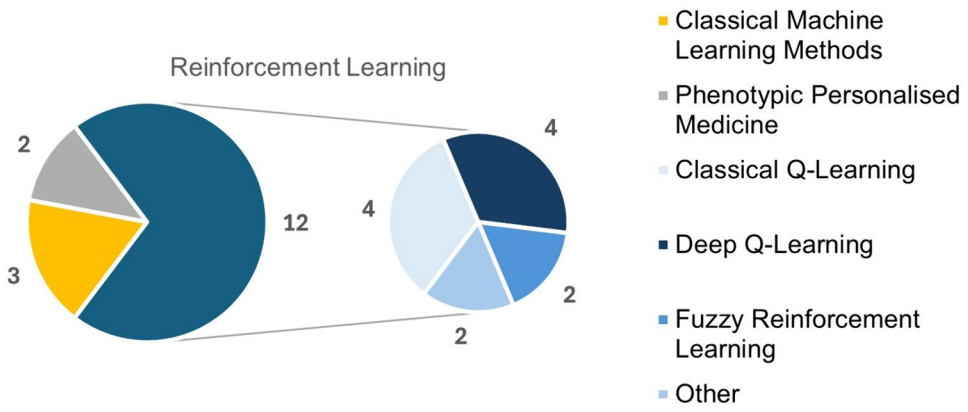



**Fig. 2** Pie chart of identified studies. The identified studies comprise 12 Reinforcement Learning studies, three Classical Machine Learning studies and two Phenotypic Personalised Medicine studies. The 12 Reinforcement Learning studies include four Classical Q-Learning, four Deep Q-Learning, two Fuzzy Reinforcement Learning and two other studies

specifically on RL strategies in different steps of cancer therapy planning [21, 22]. Additionally, a similar review summarised the application of ML to facilitate Model-Informed-Precision-Dosing (MIPD) and Therapeutic Drug Monitoring [23].

#### 3.1.1 Classical Q-Learning

In many identified studies, Q-Learning and related methods were applied for dose optimisation. Q-Learning is an off-policy RL, meaning that the target value can be computed without considering how the experience was generated [24]. Every RL algorithm strives to find a balance between explorative (novel, but possibly better action) and exploitative (action the algorithm knows will bring maximal short-term reward) actions. In Q-Learning, usually the optimal action is selected corresponding to small values of ε (probability of choosing to explore), meaning that the algorithm mostly acts upon prior knowledge (exploitation) and rarely selects actions randomly (exploration). This is referred to as an "epsilon-greedy" strategy. The Q-values, which are used to determine how good an action taken at a particular state is, are in tabular form in classical Q-Learning. They are updated based on the reward received for a state-action pair and the estimated value of the next state. By repeatedly updating the Q-values based on the observed rewards, the agent can converge to an optimal policy that maximises the cumulative reward over time. Other possibilities to explore or evaluate action space are, for example, Monte Carlo Tree Search and temporal difference.

Yazdjerdi et al. proposed a Q-Learning approach to optimise intravenous endostatin therapy for a simulated population [25]. The model aimed to reduce the tumour size under the threshold size for tumour angiogenesis (1–2 mm in diameter) [26]. A dynamic tumour growth model was implemented on virtual patients that calculated changes in tumour volume, endothelial volume and drug concentration. The Q-Learning approach was compared with other control strategies proposed in literature, using the same mathematical model of tumour growth dynamics [27, 28] and outperformed them by achieving a smaller final tumour volume with a halved and therefore more realistic drug dose in a shorter time. Similarly, with a reduced maximum dose, the total amount of drug administered, the maximum dose, and the tumour volume could all be decreased.

A similar approach was applied to optimise the temozolomide (TMZ) schedule of simulated glioblastoma multiforme patients [29]. They proposed a hybrid modelling framework, which integrated a multi-scale cellular automation model of glioblastoma growth with a RL optimising agent [30]. The model was applied to a synthetic glioblastoma patient weighing 70 kg treated with TMZ with different tumour characteristics. Compared to the classical 7/14 regimen (7 days on/7 days off), the RL model was more effective in reducing the tumour mass. However, it should be noted that the study group did not account for concomitant radiotherapy and only simulated a tumour size of 1 $mm^3$, whereas glioblastoma can grow up to 6 cm in diameter before posing a threat to a patient's life.

Most recently, Q-Learning was also applied to personalise erdafitinib protocols in patients with metastatic urothelial carcinoma [31]. Only doses available in practice could be administered and the aim was to maintain serum phosphate concentrations within a target range. A population of 141 simulated patients [32, 33] consisting of complete and partial responders was assessed at the second week of treatment, the end of the fourth treatment month and at the end of treatment. The results were compared to the U.S. Food and Drug Administration (FDA)-approved adaptive dosing protocol for erdafitinib [34]. At each timepoint, the model-based individual protocols resulted in a higher percentage of patients with serum phosphate levels within the proposed range. The RL model recommended lower starting doses than proposed in the FDA protocol and gradual dose increasing.

Padmanabhan et al. applied Q-Learning to develop an optimal controller for non-specific intravenous cancer chemotherapy drug dosing [35–37]. A non-linear four-state model [38, 39] was used to represent tumour growth. It captured the

logistic growth of the tumour, the immune response to the chemotherapy, cell proliferation and death. The aim of the controller was to optimise the dose to maximise the desired tumour shrinkage and minimise drug-induced side effects for different simulated patient groups, consisting of adults, pregnant women and multimorbid elderly patients. The resulting algorithm successfully reached the desired state for each patient group, on average after 28 days of simulation.

#### 3.1.2 Deep Q-Learning

In Deep Q-Networks and Deep Double Q-Learning, optimal response functions are represented as Deep Neural Networks with parameters (or weights) instead of a table as in classical Q-Learning. These advancements can therefore capture bigger datasets and cases. An advanced Deep Double Q-Learning can avoid the overly optimistic Q-value estimates seen in Deep Q-Learning or simple Q-Learning by separating action choice and action reward in the calculations of the Q-value, resulting in more accurate optimisation [40].

Yauney et al. applied Deep Q-Learning to guide treatment of glioma with temozolomide or a combination of procarbazine, 1,2-chloroethyl-1-nitrosurea and vincristine [41]. The Deep Q-Learning algorithm interacted with an environment of tumour growth inhibition [42] to select the appropriate dose and minimise the tumour size in 50 simulated patients for each regimen. Experiments, in which the RL agent could treat different simulated patients independently (patient-based experiments) and in which the agent had to administer the same dose to each patient (trial-based experiments) as well as experiments with fixed or variable possible doses, were conducted to compare the proposed dosing policies with clinical trial dosing regimens [42–44]. The results showed that the proposed dosing policies were generally equally effective at reducing the mean tumour diameter as clinical trial regimens. However, when the RL algorithm significantly reduced the administered dose (to 25%), the resulting tumour volume was larger than if a clinical trial regimen was applied. Yauney et al. concluded that a reward function based on reducing tumour size leads to dose regimens similar to those proposed in clinical trials, which are also guided by this goal.

In a general setting, Deep Double Q-Learning was applied to derive dosing schedules for an unspecified chemotherapy drug [45]. A total of 200 virtual trial patients was simulated using a model of breast and ovarian cancer growth in mice [46, 47]. The aim was to find an optimal dosing schedule to minimise tumour cell counts while reducing toxicity, using information on relative bone marrow density. During training, the Deep Double Q-Learning agent was unaware of patient-specific values and was only provided with average values. Drug doses and times were discretised. After enriching the model with information on relative bone marrow density, it was compared to a traditional nominal optimal controller and a nearest testing neighbour optimal controller. When tested on unknown patients, the RL agent schedules were closer to the theoretical optimum compared to the other controllers. The benefit of using RL increased with the level of parameter perturbation.

With the similar aim to maximise efficacy and reduce toxicity, Huo et al. used Multi-Objective Deep Q-Network based on Multi-Indicator Experience Replay (MIER-MO-DQN), to optimise a general chemotherapy schedule [48]. The number of effector immune cells and the administered drug dose needed to be maintained, while minimising the number of tumour cells. Two patients in good and in poor general condition were simulated, choosing a higher immune cell threshold for the latter. To quantify the value of each possible action of the model, a composite score was used, consisting of the temporal difference error, the information entropy, and the number of replays and repetitions within the optimisation. Each action was weighted to correct for divergence. The MIER-MO-DQN was compared with the conventional Deep Q-Network (DQN) approach and the Linear Weighted Sum Function-based DQN (W_DQN). The algorithms were compared in terms of changes in tumour cell, immune cell and circulating lymphocyte counts and drug concentrations. For both patients, MIER-MO-DQN performed best with a low tumour cell count, sufficiently high immune cell count and no restrictions violated.

A model-free Deep Reinforcement Learning-based method for chemotherapy drug dosing was proposed [49] and compared to a classical Q-Learning method described in Sect. 3.1.1 [35] and non-RL controllers. A non-linear pharmacological cancer model served as the environment for simulating patient data and cancer dynamics. The structural model described by Padmanabhan et al. was applied [35] and state variables and control actions were modelled continuously to avoid expert-guided discretisation [35]. The proposed Deep RL method showed a similar trend to the classical Q-Learning method [35] in several respects, but with a shortened time to reach the target state and lower drug exposure. Compared to non-RL controllers, the total administered dose was again reduced, except when compared to a state feedback control strategy.

#### 3.1.3 Fuzzy Reinforcement Learning

Fuzzy Reinforcement Learning (FRL), a fuzzy extension of Q-Learning, combines fuzzy systems as a comprehensive approximator with the principles of RL. Fuzzy logic is an approach to variable processing based on "degrees of truth" rather than binary encoding. Fuzzy Reinforcement Learning is especially useful if data, goals and constraints to the model are fuzzy in nature. In the context of dose individualisation in oncology, this could mean not only classifying

a patient as overdosed but specifying a degree of overdosing and formulating rules accordingly.

An adaptive controller based on a RL scheme with two fuzzy rules, was used to optimise an unspecified chemotherapy treatment regimen by Treesatayapun et al. [50, 51]. Four patients with different responses to drug therapy were simulated according to parameters specified in oncologic studies [52–55]. The goal of the model-free adaptive controller was to achieve complete eradication of tumour cells. The dynamics of the cell populations (normal cells, immune cells and tumour cells) was reformulated by pseudo-partial derivatives as drug administration and tumour cell population. Fuzzy rules were applied to if-then rules imposed by human knowledge according to PK and PD behaviour (for example, "if the tumour cell population is high, the drug administration must be increased"). The results showed that the proposed algorithm matched different patient needs, with more sensitive patients receiving a lower dose. Delaying treatment resulted in a marked shift in the concentration curves as expected.

Moreover, a similar approach was proposed by Alsaadi et al. [56]. In contrast to Treesatayapun et al. [50, 51], young and elderly patients were simulated with different parameter values with or without consideration of parameter uncertainty. The aims were to achieve a desired number of tumour cells ($T = 0$) for young patients and to control normal and tumour cell counts for the elderly to avoid toxicity. A fuzzy RL-based state-action-reward-state-action (SARSA) algorithm was used to control the tumour entity by chemotherapy, modelled by a Caputo-Fabrizio fractional order model [39]. Fuzzy logic was applied to enhance the controller's ability to handle uncertainty and imprecision in the system. For both patient cases, the algorithms achieved their respective goals. The proposed method was superior to simple Q-Learning in terms of efficiency, prediction error and drug dosage.

#### 3.1.4 Other Reinforcement Learning Approaches

A combined Bayesian Data Assimilation-Reinforcement Learning (DA-RL) algorithm was used to guide model-informed precision dosing of paclitaxel in simulated non-small cell lung cancer (NSCLC) patients [57]. The goal of the algorithm was to maintain patients' neutropenia grades between 1 and 3 to achieve optimal efficacy and minimal toxicity. A Monte Carlo Tree Search was used with an upper confidence bound applied to the trees. In total, 1000 patients were simulated according to values reported in a clinical study [58]. Standard dosing, PK-guided dosing, maximum-a-posteriori (MAP)-guided dosing and Bayesian data assimilation (DA) with and without RL were compared. It was shown that PK-guided dosing performed better than standard dosing, but with an increased incidence of grade 0 neutropenia as an indicator of non-efficacy. With MAP-guided dosing, the incidence of grade 4 neutropenia increased in a cumulative trend. Compared to PK-guided dosing, Bayesian DA with or without RL was able to significantly minimise the percentage of patients in neutropenia grades 0 and 4. Furthermore, the incidence of grade 0 and 4 neutropenia was significantly reduced in later cycles, highlighting the critical role of individualised uncertainty quantification. Maier et al. suggested that the small differences observed between DA with and without RL could be related to the different weighting of levels 0 and 4 in the respective reward functions [57]. They postulated that DA with RL may have great potential for long-term optimisation in a delayed-feedback environment and the integration of multiple endpoints.

On the other hand, Shiranthika et al. applied Conservative Q-Learning to optimise treatment regimens using a supervised optimal chemotherapy regimen (SOCR) approach [59]. Conservative Q-Learning is postulated to deal with possible overestimation losses better than simple Q-Learning. The algorithm was applied to 40 retrospective patients with stage 4 colon cancer receiving first-line chemotherapy based on bevacizumab or cetuximab. The reward function included changes in tumour size, patient weight and drug response, overall side effects and patient death. Proposed schedules were monitored by considering oncologists' previous treatment decisions and an adjustment factor of 0.4 was chosen to mitigate discrepancies (60% of drug doses chosen by SOCR and 40% chosen by experts). Proposed RL schedules were compared to actual prescribed schedules for six randomly selected bevacizumab patients. The RL schedules were shown to be consistent between cycles. Root mean squared error (RMSE) differences were acceptable and comparatively small. As a limitation, the reward function could be made even more accurate if it included both short-term and long-term factors according to the authors. In addition, they sometimes had to extrapolate tumour sizes and noted that weight and side effects were the only indicators of a patient's condition.

### 3.2 Methods Other Than Reinforcement Learning

In five studies, methods other than RL were used, including classical ML approaches and Phenotypic Personalised Medicine (Table 3).

#### 3.2.1 Classical ML Approaches

Aiming at predicting response to platinum-based doublet chemotherapy to optimise regimens in advanced

non-resectable NSCLC patients, Kozłowska et al. proposed a computational platform combining ML with a mechanistic mathematical model using data from 42 NSCLC patients [60]. The mathematical model was an extension of a metastatic relapse model [61] calibrated with a multivariate Gaussian mixture model estimated via an expectation-maximisation method. It included cancer cells sensitive and resistant to platinum-based chemotherapy and modelled their competition for resources. After predicting the response to chemotherapy, different schedules including applying the maximum tolerated dose and metronomic therapy, both with and without drug holidays, were compared. If competition between sensitive and resistant cells was low, there was no difference between schedules with or without drug holidays. However, if competition was assumed to be high, the best outcome occurred in schedules including drug holidays, as the number of drug-sensitive cells increased during that time.

In order to predict optimal lapatinib treatment regimens in breast cancer patients, Yu et al. applied a sequential forward-selection algorithm based on random forests for feature selection and different ML algorithms for treatment selection [62]. Data were retrospectively collected from 149 HER2(+) breast cancer patients who received either regimens including 1000 mg or 1250 mg of lapatinib. The outcome variable was the initial dose regimen of lapatinib converted to a binary variable. A sequential forward selection (SFS) algorithm based on a Random Forest was used to select the most influential features. In the next step, different algorithms (Table 3) were used for dose regimen prediction and compared for their predictive ability. The actual dose administered to the patient was used as the reference. The four most important features identified by the SFS algorithm were weight, number of prior chemotherapy treatments, number of metastases and especially the underlying treatment protocol. The most accurate regimen prediction algorithm was TabNet (Deep Neural Network for structured tabular data). Both regimens could be predicted with an accuracy of over 80%. The main limitation of the study according to the authors was the limited sample size.

Assessing exposure-effect relationships of cisplatin in head and neck cancer patients, Cauvin et al. compared different ML approaches [63]. Data were retrospectively collected from 80 patients with stage 3–4 disease. Treatment response was assessed after 12 weeks of treatment and defined according to RECIST 1.1 criteria [64]. Patients were divided into responders and non-responders. Nephrotoxicity was considered the only dose-limiting toxicity and was reported according to the CTCAE grading [65]. Pharmacokinetics was assumed to follow a three-compartment model and exposure parameters were estimated. In the next step, different ML algorithms (Supplementary Table 3) were compared to identify the exposure parameter most strongly associated with response and an optimal therapeutic range was identified by Tree-structured Parzen estimation. A theoretical dose proposal for the next chemotherapy cycle was calculated using the Kinetic Pro V1.0.3 software (IMMPS, Paris, France) by simulating expected cisplatin concentrations using PK parameters. Finally, the actual and model-guided dosing were compared in terms of compliance with the proposed therapeutic range and reported toxicity and efficacy. The most accurate algorithm for describing the relationship between exposure parameters and outcome was the Generalised Linear Model (GLM) with an accuracy of 0.71 for outcome prediction. Peak plasma concentration ($C_{max}$) was found to be most useful to guide dosing and a range was proposed. Following model-guided dosing, most patients would have been treated more adequately concerning the target range.

#### 3.2.2 Phenotypic Personalised Medicine Case Studies

CURATE.AI is an artificial intelligence platform used to correlate drug dose inputs with efficacy or toxicity outputs by parabolic equations to guide precision dosing [66]. The phenotypic map generated by the algorithm can implicitly incorporate influential mechanistic components such as disease biology, genetics or PK without explicit knowledge and the need for assumptions.

CURATE.AI has so far been used to retrospectively guide maintenance therapy in two paediatric patients with standard-risk acute lymphoblastic leukaemia on either four- or two-drug maintenance therapy [67]. It aimed at optimising treatment by maintaining absolute neutrophil counts (ANCs) and platelet counts within suitable ranges and was compared to the actual prescribed doses. The results of the case study showed that CURATE.AI resulted in better ANC ranges in both patients studied for the four- and two-drug maintenance regimens compared to the prescribed regimens. Platelet counts were within range for both strategies. In addition, the dose suggested by CURATE.AI was usually lower than the actual prescribed dose.

Furthermore, the CURATE.AI platform was used to prospectively guide the dosing of a novel bromodomain inhibitor (ZEN-3694) and enzalutamide in a Phase 1b/2a safety and tolerability trial [68]. The drugs were administered to an enzalutamide-naïve 82-year-old male patient with metastatic castration-resistant prostate cancer and alleviated prostate-specific antigen (PSA) levels despite previous abiraterone chemotherapy. For the first 6 months of treatment, the dose was guided by a physician only; the following 6 months, CURATE.AI recommendations were considered. Patient PSA levels were used as the primary marker of clinical activity to guide dosing. During the physician-guided phase, the patient's PSA level and the size of the target lesion already decreased significantly after dose reductions of both drugs.

At the end of CURATE.AI-guided dosing, the patient's PSA level dropped even further. Overall, the implementation of CURATE.AI resulted in a durable response using a reduced dose of ZEN-3694, although ZEN-3694 was found to be a key modulator of treatment response. However, a major limitation of both case studies on CURATE.AI is the limited number of patients included.

## 4 Discussion

In general, 17 studies on the use of ML methods for antitumour therapy optimisation could be identified in the literature search, most of which applied RL. Within the identified RL studies, epsilon-greedy strategies have been used frequently. While this strategy performs particularly well in the short term, less greedy strategies may be interesting in the long term, which stresses more on exploration [57]. More recent studies have used more advanced Q-Learning strategies in which optimal response functions are represented as Deep Neural Networks instead of a table, making them potentially more suitable for complex scenarios in oncology. In the publications described, classical Q-Learning was compared with a FRL-based control method (SARSA algorithm) [56] and a DQN [49] with the novel methods showing superiority in terms of maximising the reward and minimising therapy duration. In addition, when a Multi-Objective Deep Q-Network was compared with a conventional DQN approach, combining multiple objectives proved to be beneficial [48]. Generally speaking, algorithms performed especially well when they considered both efficacy and toxicity within their reward functions [31, 48, 57]. Moreover, it should be emphasised that the acceptance of using a ML method to optimise cancer treatment schedules may increase if patient-relevant parameters and expert opinions are considered [59].

Other than RL, a few identified studies compared various classical methods in terms of their predictive abilities. In these studies, TabNet, which is a Deep Neural Network designed for tabular data [62], and a GLM [63] demonstrated superior performance compared to other methods for their respective tasks. However, since the studies vastly differ in their therapy context and the tasks the algorithms were applied for, they are difficult to compare. One study combined ML methods with mathematical models, which led to promising results [60]. Furthermore, the AI platform CURATE.AI, which was used to optimise combination therapy simultaneously, appears to be interesting for clinical practice, as anticancer therapy is usually a combination of drugs [67, 68]. However, it has only been tested on very few cancer patients so far.

Although model and environment parameters were often reported for the RL algorithms, resulting doses were more often reported for the other algorithms. In future research, resulting doses and accuracy measures (if applicable) should be reported along with important hyperparameters or optimally the whole code. Due to the significantly different algorithm structures, the performance of RL and non-RL methods cannot be compared. Furthermore, if direct comparisons were not made for the specific example, it is not possible to directly compare different RL or non-RL methods among each other, respectively. However, it can be stated that RL algorithms are applicable for more complex tasks and can be used for dynamic dose adjustments in ongoing or retrospective therapies. Therefore, research might focus more on RL in the future and ways to apply it in a clinical setting.

While some studies guided dosing of individual antitumour drugs, other studies dealt with the method itself and are yet to be made applicable in pracitice. However, clinical utility is yet to be proven for all algorithms described in this review. Future work should focus on combining efficacy and toxicity measures and patient-relevant parameters within the reward, as this approach should yield the most promising, acceptable and clinically relevant results. In clinical routine, where the focus is not only on survival but also on maintaining quality of life in chronic disease, there is no point in proposing regimens that improve efficacy but do not also ensure that toxicity is not worsened. Moreover, the proposed dosing regimens must be practical and avoid irregularities (therapy at inappropriate times and dates and inadequate doses). In addition, to be closer to practice, algorithms need to incorporate concentration-response relationships that have been studied in humans and not just incorporate cellular interactions. To add to their value, these algorithms should preferably be able to optimise combination chemotherapy regimens, as anticancer drugs are usually used jointly. Such algorithms should then be compared with Therapeutic Drug Monitoring or other common dosing strategies to investigate their clinical utility in terms of maximising efficacy (e.g., by measuring tumour growth or assessing the incidence of metastases) and avoiding toxicity. Adverse events should not only be reported by the clinical staff based on the CTCAE criteria [65] but also by the patients themselves, e.g., using the patient-reported outcome version of the CTCAE [69].

If proven clinically useful, RL agents could potentially change the dosing paradigms of both classical and targeted anticancer agents, and pose an alternative to the still common fixed-dose approach [70]. In classical chemotherapy, they could be used to minimise the number of treatment cycles required, particularly to avoid long-term toxicity and secondary cancers [71]. In addition, if the algorithms can propose schedules with minimal acute adverse events, there would potentially be fewer treatment interruptions and the need for supplementary therapy. In the case of oral targeted

anticancer therapy, optimised schedules could also reduce the incidence of adverse events and improve patient adherence [72]. Due to the high cost of oral targeted anticancer drugs, rising faster than inflation [73], another useful goal may be to minimise treatment costs by testing whether less frequent dosing can achieve comparable outcomes. This could enable the use of oral targeted anticancer therapy in less wealthy countries and patient groups, thereby increasing fairness and equity in cancer treatment. In a recently published review on Model-Informed Reinforcement Learning for precision dosing of different drugs, it was emphasised that clinical knowledge and constraints need to be taken into account to obtain useful adaptive dosing strategies and that algorithms need to be tested prospectively and not only in silico [74]. Additionally, the authors state that similarly to other methods, RL dose optimisation may be suboptimal or inefficient for patient cohorts with high inter-individual variability. In such cases, implementing the RL-based optimisation in a Bayesian fashion may be helpful [57, 74].

Overall, it is worth noting that in our literature search we were able to identify further preprints in this area which could not be included yet for lack of peer review. This highlights the importance of this area of research and that many more studies will be published in the near future. However, it also calls attention to the main limitation of this scoping review: The number of studies published on this topic is constantly increasing and there might be new eligible studies which are not included upon publication. Additionally, only studies reporting results were included in this review, while there might have been further interesting methodological publications (i.e., conference presentations) which did not yet present final results. Direct comparisons of the methods were mostly impossible as different methods were applied in different therapy contexts or general settings. In addition, because a scoping review does not assess the quality of the evidence, it cannot assess the implications for practice or policy.

## 5 Conclusions

In summary, this review provides a comprehensive overview on ML methods that were recently used and evaluated to optimise cancer treatment dosing. Twelve of the 17 included studies used RL methods, including Classical, Deep, Deep Double and Conservative Q-Learning and Fuzzy RL. In many cases, a tumour growth model was proposed to describe changes in the number of tumour cells, immune cells, healthy cells and drug concentration during cancer therapy. In these cases, the algorithm was mostly rewarded for minimising tumour size and for sparing healthy cells. Some trials included toxicity endpoints and patient-relevant parameters in addition to efficacy endpoints in their approach, such as biomarker levels, changes in tumour size, side effects and patient death. In most cases, epsilon greedy strategies have been used. Furthermore, classical ML methods were compared in terms of their performance, ML and mathematical modelling have been combined and an artificial intelligence platform has been used to guide dosing prospectively and retrospectively, albeit only in very few patients. Future studies will probably continue exploring advanced Q-Learning for dose optimisation and consider drug efficacy, toxicity and patient-relevant parameters within the reward. Overall, ML methods have a great potential for maximising efficacy and minimising toxicity by dose optimisation when compared to standard protocols even by model-free predictions.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s40262-024-01409-9.

**Acknowledgements** We thank Ana Socorro Rodríguez-Báez for proofreading and editing this review.

## Declarations

**Funding** Open Access funding enabled and organized by Projekt DEAL. ET received funding by the German Federal Ministry of Education and Research within the project "BNTrAinee" (funding code 16DHBK1022). HF received grants from UCB and AbbVie. The funders were not directly involved in this work.

**Conflict of interest** The authors disclose no conflicts of interest.

**Data, material and code availability** Not applicable.

**Ethics approval** Not applicable.

**Author contributions** UJ designed the research; OT carried out the literature search; OT and ME screened the identified publications, DV adjudicated in case of disagreement; OT and LMK extracted the relevant data; OT, LMK, ME, MV, DV, ET, HF, JH and UJ wrote and refined the manuscript.

**Use of AI** The automated tool Rayyan was used for abstract and full text screening. The tool DeepL Write was used by OT to refine her text.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

## References


1. Hu PJH, Wei CP, Cheng TH, Chen JX. Predicting adequacy of vancomycin regimens: a learning-based classification approach to improving clinical decision making. Decis Support Syst. 2007;43:1226–41.
2. Imai S, Takekuma Y, Miyai T, Sugawara M. A new algorithm optimized for initial dose settings of vancomycin using machine learning. Biol Pharm Bull. 2020;43:188–93.
3. Tang J, Liu R, Zhang Y-L, et al. Application of machine-learning models to predict tacrolimus stable dose in renal transplant recipients. Sci Rep. 2017;7:42192.
4. Lu J, Deng K, Zhang X, et al. Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens. iScience. 2021;24:102804.
5. You Dubout W. An algorithmic approach to personalized drug concentration predictions. Lausanne: EPFL; 2014.
6. Stankevičiūtė K, Woillard JB, Peck RW, et al. Bridging the worlds of pharmacometrics and machine learning. Clin Pharmacokinet. 2023;62:1551–65.
7. Chen S, Peng Y, Qin A, et al. MR-based synthetic CT image for intensity-modulated proton treatment planning of nasopharyngeal carcinoma patients. Acta Oncol. 2022;61:1417–24.
8. The 2022 AAPM Annual Meeting Program. Med Phys. 2022;49:e113–e982.
9. Zhao J, Chen Z, Wang J, et al. MV CBCT-based synthetic CT generation using a Deep Learning method for rectal cancer adaptive radiotherapy. Front Oncol. 2021;11: 655325.
10. Men K, Zhang T, Chen X, et al. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Phys Med. 2018;50:13–9.
11. Kawata Y, Arimura H, Ikushima K, et al. Impact of pixel-based machine-learning techniques on automated frameworks for delineation of gross tumor volume regions for stereotactic body radiation therapy. Phys Med. 2017;42:141–9.
12. Kawula M, Purice D, Li M, et al. Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer. Radiat Oncol. 2022;17:21.
13. Osman AFI, Tamam NM. Attention-aware 3D U-Net convolutional neural network for knowledge-based planning 3D dose distribution prediction of head-and-neck cancer. J Appl Clin Med Phys. 2022;23: e13630.
14. Frederick A, Roumeliotis M, Grendarova P, Quirk S. Performance of a knowledge-based planning model for optimizing intensity-modulated radiotherapy plans for partial breast irradiation. J Appl Clin Med Phys. 2022;23: e13506.
15. de Dios NR, Moñino AM, Liu C, et al. Machine learning-based automated planning for hippocampal avoidance prophylactic cranial irradiation. Clin Transl Oncol. 2023;25:503–9.
16. Peters M, Godfrey C, McInerney P, Munn Z, Tricco A, Khalil H. Chapter 11: scoping reviews (2020 version). 2020. In: JBI manual for evidence synthesis. JBI; 2020. https://synthesismanual.jbi.global. Accessed 11 Dec 2023.
17. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169:467–73.
18. Rayyan-AI powered tool for systematic literature reviews. https://www.rayyan.ai/.
19. Teplytska O. Review protocol; 2023. https://osf.io/qm3yr/.
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372: n71.
21. Ribba B, Kaloshi G, Peyre M, et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. Clin Cancer Res. 2012;18:5071–80.
22. Yang CY, Shiranthika C, Wang CY, et al. Reinforcement learning strategies in cancer chemotherapy treatments: a review. Comput Methods Progr Biomed. 2023;229: 107280.
23. Poweleit EA, Vinks AA, Mizuno T. Artificial intelligence and machine learning approaches to facilitate therapeutic drug management and model-informed precision dosing. Ther Drug Monit. 2023;45:143–50.
24. Watkins CJ, Dayan P. Technical note: Q-learning. Mach Learn. 1992;8:279–92.
25. Yazdjerdi P, Meskin N, Al-Naemi M, et al. Reinforcement learning-based control of tumor growth under anti-angiogenic therapy. Comput Methods Progr Biomed. 2019;173:15–26.
26. Drexler DA, Sápi J, Szeles A, et al. Flat control of tumor growth with angiogenic inhibition. In: 7th IEEE International Symposium 2012. p. 179–83.
27. Sápi J, Drexler DA, Harmati I, et al. Linear state-feedback control synthesis of tumor growth control in antiangiogenic therapy. In: 10th IEEE International Symposium 2014. p. 143–8.
28. Drexler DA, Kovács L, Sápi J, et al. Model-based analysis and synthesis of tumor growth under angiogenic inhibition: a case study*. IFAC Proc Vol. 2011;44:3753–8.
29. Ebrahimi Zade A, Shahabi Haghighi S, Soltani M. Reinforcement learning for optimal scheduling of glioblastoma treatment with temozolomide. Comput Methods Progr Biomed. 2020;193: 105443.
30. Stamatakos GS, Antipas VP, Uzunoglu NK. A spatiotemporal, patient individualized simulation model of solid tumor response to chemotherapy in vivo: the paradigm of glioblastoma multiforme treated by temozolomide. IEEE Trans Biomed Eng. 2006;53:1467–77.
31. de Carlo A, Tosca EM, Fantozzi M, Magni P. Reinforcement learning and PK-PD models integration to personalize the adaptive dosing protocol of erdafitinib in patients with metastatic urothelial carcinoma. Clin Pharmacol Ther. 2024.
32. Dosne AG, 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. J Clin Pharmacol. 2020;60:515–27.
33. Dosne AG, Valade E, Stuyckens K, et al. Erdafitinib's effect on serum phosphate justifies its pharmacodynamically guided dosing in patients with cancer. CPT Pharmacometr Syst Pharmacol. 2022;11:569–80.
34. Janssen Pharmaceutical Companies. BALVERSA (erdafitinib) tablets, for oral use initial U.S. approval: 2019. 2019.
35. Padmanabhan R, Meskin N, Haddad WM. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. Math Biosci. 2017;293:11–20.
36. Padmanabhan R, Meskin N, Haddad WM. Learning-based control of cancer chemotherapy treatment. IFAC-PapersOnLine. 2017;50:15127–32.
37. Padmanabhan R, Meskin N, Haddad WM. 9—Reinforcement learning-based control of drug dosing with applications to anesthesia and cancer therapy. In: Control applications for biomedical engineering systems. Academic Press: New York; 2020. p. 251–97.
38. Batmani Y, Khaloozadeh H. Optimal chemotherapy in cancer treatment: state dependent Riccati equation control and extended Kalman filter. Optim Control Appl Methods. 2013;34:562–77.
39. de Pillis L, Radunskaya A. The dynamics of an optimally controlled tumor model: a case study. Math Comput Model. 2003;37:1221–44.

40. Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 30th edn. 2016.
41. Yauney G, Shah P. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. In: Proceedings of the 3rd Machine Learning for Healthcare Conference. Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection. PMLR; 2018. p. 161–226.
42. Ribba B, Dudal S, Lavé T, Peck RW. Model-informed artificial intelligence: reinforcement learning for precision dosing. Clin Pharmacol Ther. 2020;107:853–7.
43. Ricard D, Kaloshi G, Amiel-Benouaich A, et al. Dynamic history of low-grade gliomas before and after temozolomide treatment. Ann Neurol. 2007;61:484–90.
44. Peyre M, Cartalat-Carel S, Meyronet D, et al. Prolonged response without prolonged chemotherapy: a lesson from PCV chemotherapy in low-grade gliomas. Neuro Oncol. 2010;12:1078–82.
45. Eastman B, Przedborski M, Kohandel M. Reinforcement learning derived chemotherapeutic schedules for robust patient-specific therapy. Sci Rep. 2021;11:17882.
46. Panetta JC. A mathematical model of breast and ovarian cancer treated with paclitaxel. Math Biosci. 1997;146:89–113.
47. Panetta JC, Adam J. A mathematical model of cycle-specific chemotherapy. Math Comput Model. 1995;22:67–82.
48. Huo L, Tang Y. Multi-objective deep reinforcement learning for personalized dose optimization based on multi-indicator experience replay. Appl Sci. 2023;13:325.
49. Mashayekhi H, Nazari M, Jafarinejad F, Meskin N. Deep reinforcement learning-based control of chemo-drug dose in cancer treatment. Comput Methods Progr Biomed. 2024;243: 107884.
50. Treesatayapun C, Muñoz-Vázquez AJ. Optimal drug-dosing of cancer dynamics with fuzzy reinforcement learning and discontinuous reward function. Eng Appl Artif Intell. 2023;120: 105851.
51. Treesatayapun C, Muñoz-Vázquez AJ, Suyaroj N. Reinforcement learning optimal control with semi-continuous reward function and fuzzy-rules networks for drug administration of cancer treatment. Soft Comput. 2023;27:17347–56.
52. Ekpenyong ME, Etebong PI, Jackson TC, Udofa EM. Modelling drugs interaction in treatment-experienced patients on antiretroviral therapy. Soft Comput. 2020;24:17349–64.
53. Sharifi M, Moradi H. Nonlinear composite adaptive control of cancer chemotherapy with online identification of uncertain parameters. Biomed Signal Process Control. 2019;49:360–74.
54. Rihan FA, Velmurugan G. Dynamics of fractional-order delay differential model for tumor-immune system. Chaos Solitons Fractals. 2020;132: 109592.
55. Babaei N, Salamci MU. Personalized drug administration for cancer treatment using model reference adaptive control. J Theor Biol. 2015;371:24–44.
56. Alsaadi FE, Yasami A, Volos C, et al. A new fuzzy reinforcement learning method for effective chemotherapy. Mathematics. 2023;11:477.
57. Maier C, Hartung N, Kloft C, et al. Reinforcement learning and Bayesian data assimilation for model-informed precision dosing in oncology. CPT Pharmacometr Syst Pharmacol. 2021;10:241–54.
58. 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.
59. Shiranthika C, Chen K-W, Wang C-Y, et al. Supervised optimal chemotherapy regimen based on offline reinforcement learning. IEEE J Biomed Health Inform. 2022;26:4763–72.
60. Kozłowska E, Suwiński R, Giglok M, et al. Mathematical model predicts response to chemotherapy in advanced non-resectable non-small cell lung cancer patients treated with platinum-based doublet. PLoS Comput Biol. 2020;16: e1008234.
61. Nicolò C, Périer C, Prague M, et al. Machine learning and mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. JCO Clin Cancer Inform. 2020;4:259–74.
62. Yu Z, Ye X, Liu H, et al. Predicting lapatinib dose regimen using machine learning and deep learning techniques based on a real-world study. Front Oncol. 2022;12: 893966.
63. Cauvin C, Bourguignon L, Carriat L, et al. Machine-learning exploration of exposure–effect relationships of cisplatin in head and neck cancer patients. Pharmaceutics. 2022;14:2509.
64. RECIST 1.1 criteria. https://recist.eortc.org/recist-1-1-2/. Accessed 03 May 2024.
65. Common Terminology Criteria for Adverse Events (CTCAE) v5.0. https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm#ctc_60. Accessed 27 Dec 2023.
66. Blasiak A, Khong J, Kee T. CURATE.AI: optimizing personalized medicine with artificial intelligence. SLAS Technol. 2020;25:95–105.
67. Lee DK, Chang VY, Kee T, et al. Optimizing combination therapy for acute lymphoblastic leukemia using a phenotypic personalized medicine digital health platform: retrospective optimization individualizes patient regimens to maximize efficacy and safety. SLAS Technol. 2017;22:276–88.
68. Pantuck AJ, Lee D-K, Kee T, et al. Modulating BET bromodomain inhibitor ZEN-3694 and enzalutamide combination dosing in a metastatic prostate cancer patient using CURATE.AI, an artificial intelligence platform. Adv Ther. 2018;1:1800104.
69. Overview of the PRO-CTCAE. https://healthcaredelivery.cancer.gov/pro-ctcae/overview.html. Accessed 26 July 2024.
70. Mueller-Schoell A, Groenland SL, Scherf-Clavel O, et al. Therapeutic drug monitoring of oral targeted antineoplastic drugs. Eur J Clin Pharmacol. 2021;77:441–64.
71. Demoor-Goldschmidt C, de Vathaire F. Review of risk factors of secondary cancers among cancer survivors. Br J Radiol. 2019;92:20180390.
72. Cheung WY. Difficult to swallow: issues affecting optimal adherence to oral anticancer agents. Am Soc Clin Oncol Educ Book. 2013;33:265–70.
73. Seiger K, Mostaghimi A, Silk AW, et al. Association of rising cost and use of oral anticancer drugs with Medicare part D spending from 2013 through 2017. JAMA Oncol. 2020;6:154–6.
74. Tosca EM, de Carlo A, Ronchi D, Magni P. Model-informed reinforcement learning for enabling precision dosing via adaptive dosing. Clin Pharmacol Ther. 2024. (**Online ahead of print**).

## Authors and Affiliations

**Olga Teplytska[1] · Moritz Ernst[2] · Luca Marie Koltermann[1] · Diego Valderrama[3] · Elena Trunz[4] · Marc Vaisband[5,6,7] · Jan Hasenauer[5,6] · Holger Fröhlich[3,8] · Ulrich Jaehde[1]**

✉ Ulrich Jaehde
u.jaehde@uni-bonn.de

1 Department of Clinical Pharmacy, Institute of Pharmacy, University of Bonn, An der Immenburg 4, 53121 Bonn, Germany

2 Faculty of Medicine and University Hospital Cologne, Institute of Public Health, University of Cologne, Cologne, Germany

3 Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

4 Institute of Computer Science II, Visual Computing, University of Bonn, Bonn, Germany

5 Hausdorff Center for Mathematics, University of Bonn, Bonn, Germany

6 Institute of Life & Medical Sciences (LIMES), University of Bonn, Bonn, Germany

7 Department of Internal Medicine III with Haematology, Medical Oncology, Haemostaseology, Infectiology and Rheumatology, Oncologic Center, Salzburg Cancer Research Institute-Laboratory for Immunological and Molecular Cancer Research (SCRI-LIMCR), Paracelsus Medical University, Cancer Cluster Salzburg, Salzburg, Austria

8 Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn, Germany