cancers

MDPI

Review

Reinforcement Learning for Precision Oncology

Jan-Niklas Eckardt [1,*], Karsten Wendt [2], Martin Bornhäuser [1,3,4] and Jan Moritz Middeke [1]

[1] Department of Internal Medicine I, University Hospital Carl Gustav Carus, 01307 Dresden, Germany; Martin.Bornhaeuser@uniklinikum-dresden.de (M.B.); janmoritz.middeke@uniklinikum-dresden.de (J.M.M.)
[2] Institute of Software and Multimedia Technology, Technical University Dresden, 01069 Dresden, Germany; karsten.wendt@tu-dresden.de
[3] German Consortium for Translational Cancer Research, 69120 Heidelberg, Germany
[4] National Center for Tumor Diseases, 01307 Dresden, Germany
* Correspondence: Jan-Niklas.Eckardt@uniklinikum-dresden.de

Simple Summary: The accelerating merger of information technology and cancer research heralds the advent of novel methods and models for clinical decision making in oncology. Reinforcement learning—as one of the major subspecialties in machine learning—holds the potential for the development of high-performance decision support tools. However, many recent studies of reinforcement learning in oncology suffer from common shortcomings and pitfalls that need to be addressed for the development of safe, interpretable and reliable algorithms for future clinical practice.

Abstract: Precision oncology is grounded in the increasing understanding of genetic and molecular mechanisms that underly malignant disease and offer different treatment pathways for the individual patient. The growing complexity of medical data has led to the implementation of machine learning techniques that are vastly applied for risk assessment and outcome prediction using either supervised or unsupervised learning. Still largely overlooked is reinforcement learning (RL) that addresses sequential tasks by exploring the underlying dynamics of an environment and shaping it by taking actions in order to maximize cumulative rewards over time, thereby achieving optimal long-term outcomes. Recent breakthroughs in RL demonstrated remarkable results in gameplay and autonomous driving, often achieving human-like or even superhuman performance. While this type of machine learning holds the potential to become a helpful decision support tool, it comes with a set of distinctive challenges that need to be addressed to ensure applicability, validity and safety. In this review, we highlight recent advances of RL focusing on studies in oncology and point out current challenges and pitfalls that need to be accounted for in future studies in order to successfully develop RL-based decision support systems for precision oncology.

Keywords: precision oncology; reinforcement learning; artificial intelligence; machine learning; dose adjustment; chemotherapy; radiotherapy

Citation: Eckardt, J.-N.; Wendt, K.; Bornhäuser, M.; Middeke, J.M. Reinforcement Learning for Precision Oncology. *Cancers* **2021**, *13*, 4624. https://doi.org/10.3390/cancers13184624

Academic Editors: Andreas Stadlbauer, Anke Meyer-Baese and Max Zimmermann

Received: 20 August 2021
Accepted: 13 September 2021
Published: 15 September 2021

1. Introduction

Recent advances both in terms of generating an ever-growing body of medical data and the increasing computational capacity to organize such data herald an accelerating merger of information technology and the medical domain. At the intersection of increasingly more complex medical data and computational analysis, machine learning (ML) gains a foothold driven by recent developments both in hardware components and accessible software technologies [1,2]. In general, machine learning encompasses three fundamental methodologies (Figure 1): supervised learning (SL), unsupervised learning (UL) and reinforcement learning (RL) [3].

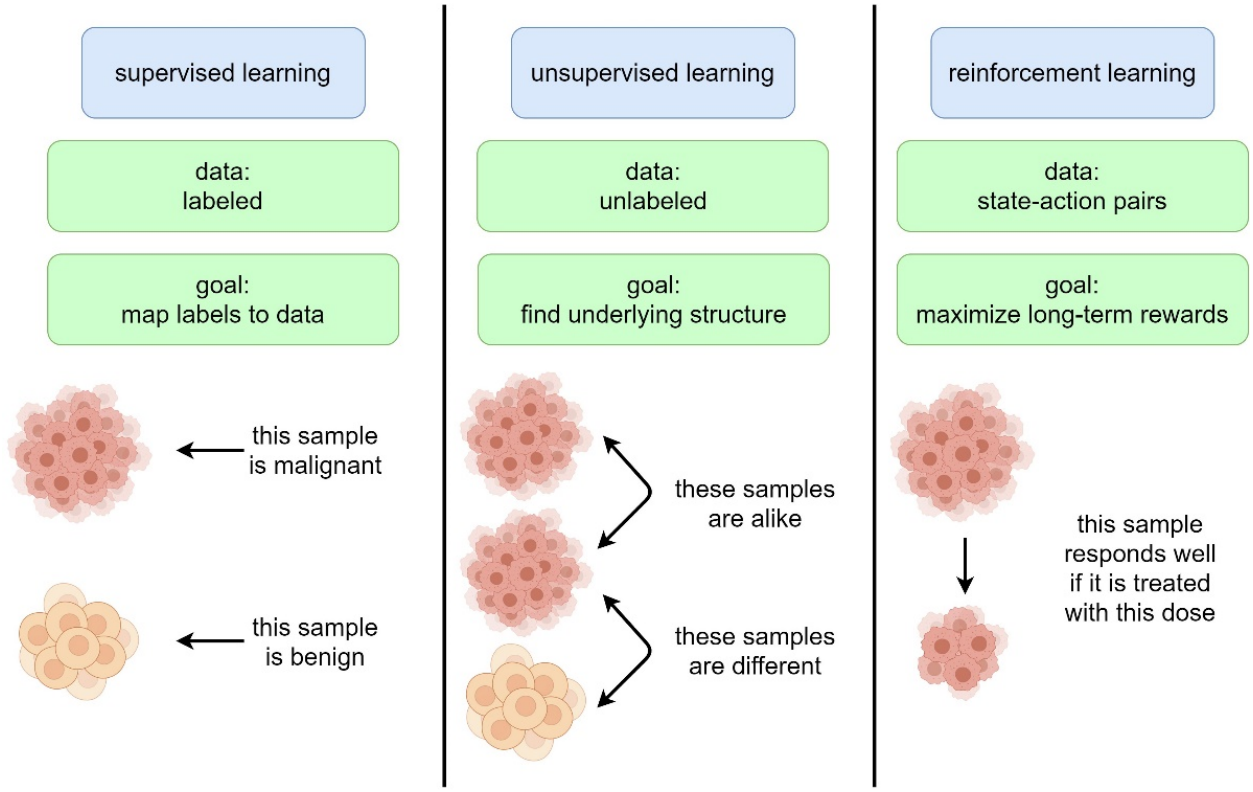

Figure 1. Main differences between machine learning techniques.

In supervised learning, an algorithm is trained on a set of previously labeled data and learns features to map labels to unlabeled data of a test set. Ideally, it can then recognize and label real-world data, for example, for class prediction in histology, where it may distinguish between benign and malignant tissue. Another example is the detection of breast cancer in radiology, where pre-labeled images (benign/malignant based on histology) can be used to train an algorithm to spot malignancies and guide subsequent treatment planning [4–9]. In unsupervised learning, the data are unlabeled and clustered based on similarities and differences. This can, for example, be used to identify groups of patients at risk using genomics where different genetic clusters of a disease may have either favorable or unfavorable outcomes [10,11]. Both these methods are broadly applied to (often retrospectively) medical datasets and are utilized for diagnosis, risk stratification, genomic clustering, outcome prediction, relapse monitoring and treatment response prediction [12]. However, clinical practice is dynamic, and the question of how well algorithms that are exclusively trained on retrospective data perform in a prospective real-world setting remains unanswered in most cases. To address the challenge of a non-stationary clinical environment with changing conditions and stimuli, RL bears the potential to develop novel methods for data-driven computer-guided decision support systems. RL learns to select different actions according to different environmental states in order to maximize long-term rewards. This may be used for dynamic treatment regimens where doses are selected according to tumor and patient biology, treatment response and adverse events to tailor a treatment strategy that fits the individual patient.

In recent years, RL has rapidly evolved, demonstrating unprecedented success and often achieving human-level or superhuman-level performance in, for example, gameplay of complex board games such as chess, Shogi and Go [13–15], video games [16–19] and the field of autonomous driving [20].

Precision medicine aims at tailoring therapy and dosing to the individual patient based on individual intrinsic factors such as patient and disease biology that may affect the response to therapy, risk of treatment failure or relapse and prognosis [21]. Consequently,

interventions can be adjusted to the individual patient or patient groups that may respond more favorably while, at the same time, reducing the risk of adverse events in patients who are unlikely to benefit. Both SL and UL currently receive the most attention as they offer insight into disease prognostication as well as treatment response using retrospective data. However, the dynamic situation both the individual patient and clinician find themselves in during oncologic treatment is not well captured by both SL and UL. The sequential foundation of RL provides a more suitable approach to capture the dynamics of oncologic therapy in a real-world (prospective) setting where both patient and environmental variables may change over the course of treatment.

In this review, we aim to provide a general understanding of the foundations of RL for the clinical oncologist, highlight previous studies of RL in oncology and outline potential pitfalls and considerations for future studies in this novel subfield at the intersection between healthcare and ML.

2. Overview of Reinforcement Learning

In this subsection, we provide a general overview of the concepts of reinforcement learning. We aim to inform the reader of the fundamental assumptions of RL, important terminology (Table 1) and different variations of RL methodologies (Table 2). For a more in-depth outline, we refer the interested reader to the detailed explanations provided by Sutton and Barto [22].

In RL, an agent interacts with its environment over time by selecting actions depending on the observed states of the environment while following a policy in order to maximize a cumulative reward (Figure 2) [22]. At each time point t, the agent observes a state st out of a pool of possible states S and selects an action from a pool of possible actions A following its policy $\pi(at \mid st)$. For its choice of action according to the observed states of the environment, the agent receives a reward rt according to a reward function R and subsequently transitions to the next state st + 1 according to a transition function T. Finally, the return the agent receives is the accumulated reward discounted by the discount factor $\gamma \in (0, 1]$ [23].

Table 1. Terminology of reinforcement learning.

Term	Symbol	Description
Reinforcement Learning	RL	operates in a simulated environment with distinct behavior to receive rewards
Environment	E	consumes actions to produce rewards for an agent; based on a model/simulation/observations/data
Agent		RL decision instance, performing actions to change states
Action	a	performed by an agent to change to another state, i.e., interact with the environment
State	s	abstract relation of the agent to the environment, starting and end point of an action
Reward	r	gain for an action of the access of a state
Reward Function	R	entirety of all rewards for actions/states
Cumulative Reward	CR	aggregated rewards of subsequent actions/states; should be maximized as the learning/optimization objective
Policy	π	defines an action for each state; result of learning/optimization process

Table 2. Variants and methodologies of reinforcement learning.

Aspect	Variant	Description	Pro	Contra
Environment	Model-Based	distinct rule-based/simulation-based feedback for the agent	covers corner cases, potentially high feedback quality	complex to set up
	Model-Free	data-based (observation/retro-perspective) feedback	easy to set up, no abstraction	no corner cases, potentially low feedback quality
Reward	V (State-Based)	rewards when accessing a state (relation to E)	fewer states, easy to model	more abstraction, static (less intuitive) view
	Q (Action-Based)	rewards when executing an action (changing E)	more actions, fewer abstraction, extensive to model	more actions, dynamic (intuitive) view
	Concluding Learning	rewards when finalizing a sequence of decisions	long term-oriented, aims for global objectives	provides no local guidance, complex evaluation
	Temporal Difference Learning	rewards after each decision	provides no local guidance, easy evaluation	short term-oriented, aims for local objectives
Access	Online	access of the agent to the E in a (restricted) stream-based way	less information to process for the agent, smaller solution space	potentially non-optimal solutions (policy)
	Batch-Based	access of the agent to the entire environment E	globally optimized solutions (policy)	more information to process, large solution space
Dynamics	Static Reward Function	each piece of feedback from the E is encoded in states, resulting in constant rewards	easier E, smaller solution space	potentially coarse-grained decisions/optimization
	Dynamic Reward Function	feedback from E is encoded in attributes, resulting in variable rewards	potentially fine-grained decisions/optimization	complex E, large solution space
	Markov Assumption	no influence from previous decisions	smaller solution space	potentially insufficient decision impact modeling
	No Markov Assumption	decision history has influence on rewards	complex decision modeling	large solution space
Representation	Table-/Map-Based	simple state transition/action modeling	easy to create, transparent	complex to maintain and show, grows exponentially with number of states
	Graph-Based	intuitive state machine modeling	easy to maintain, transparent, scales with number of states	complex to create and show
	Deep Neural Net	DL-based modeling	easy to create, scales with number of states	low transparency, complex to show

For example, an RL agent could be presented with multimodal patient data, e.g., demographics, laboratory values, tumor burden and therapy-associated toxicities, that represent the environment. For every iteration, the agent then selects an action, for example, a dose adjustment on a linear scale from 0 to 100%, given the state of the environment. This action will result in an alteration of the environment, i.e., of the patient's condition and the

data associated with it, resulting in a reward or a penalty for the agent based on whether or not the chosen action led to a favorable outcome for the patient. In that sense, the agent can abstract a policy either from rewards or state–action pairs that drives action selection, for example, the agent may learn that increasing doses of chemotherapy are associated with an increased anti-tumor effect, but also increased toxicity.

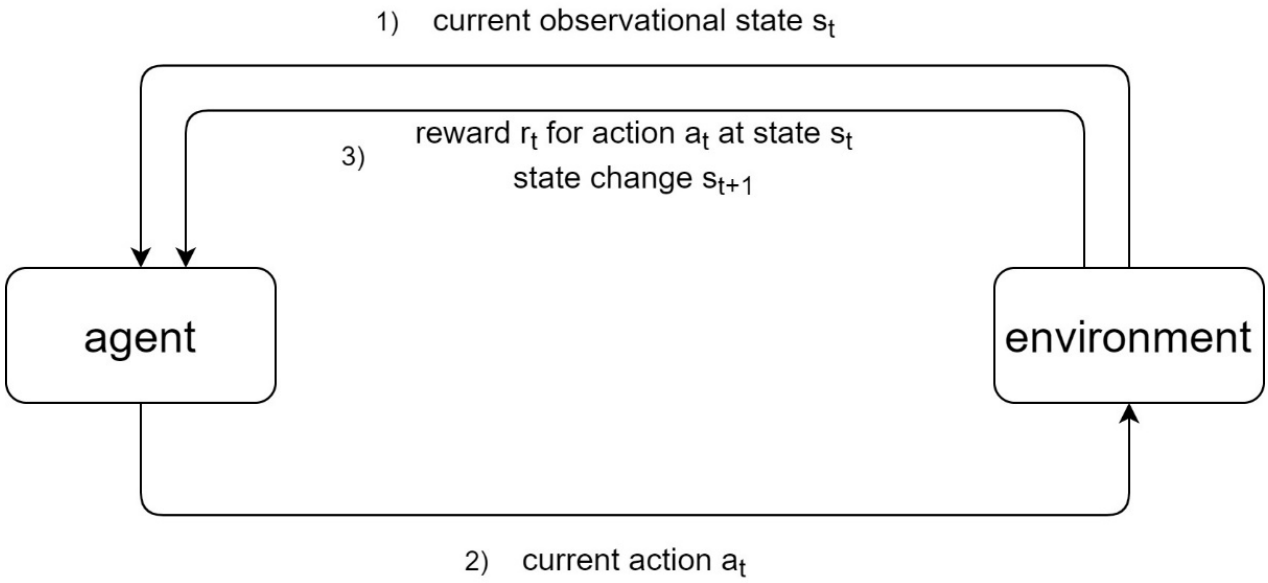

Figure 2. Concept of reinforcement learning. An agent receives current state observations from an environment (1) and, in response, selects an action according to its policy (2). For this action, the agent receives a reward based on a reward function and the state of the environment changes (3). The agent's goal is to maximize long-term rewards and achieve the optimum possible return.

Hence, the agent ultimately aims at maximizing the long-term return from each state–action pair instead of short-term rewards by selecting the appropriate actions at each given state. If the problem setup enables an underlying model to be determined or learned from experience, for example, in a game setting with clearly defined rules and transitions, learning is referred to as model-based. If the model for state transition and reward is unknown, the agent learns directly from experience using a trial-and-error approach, and learning is referred to as model-free. Regarding nomenclature, if the underlying model is simulated while the data stem from a real-world cohort, the setup of the experiment is referred to as in virtuo [24]. However, if the data are simulated as well as the model, the experiment is referred to as in silico. In virtuo experiments are frequently performed when a retrospective patient cohort is available as a data source, while in silico experiments often require modeling of plausible real-world-like data, for example, the behavior of simulated cancer cells under the influence of chemotherapy. When the agent trains on data presented in a sequential manner, learning is referred to as online, while if all data are presented at the same time (i.e., in retrospective setups), learning is referred to as offline or batch mode [23]. If the entirety of the available action and observation space is known, a future state only depends on the current state and action (Markov criterion). This problem can be described as a Markov decision process (MDP) by a tuple of (S, A, R, T, γ) [22]. Receiving the maximum possible return in an MDP environment is achieved by optimizing the agent's policy. For each policy π, a value function $V\pi$ can be determined which predicts the expected reward the agent will accumulate when acting according to policy π in a state s [22]. As an alternative to the state value function, an action value function $Q\pi(s \mid a)$ can be determined that predicts the reward based on the agent taking a specific action a in a state s [22]. Both $V\pi$ and $Q\pi$ can be expressed with the Bellman equation [25]. While both the value and policy iteration update all value states for each iteration, the temporal difference updates single state

values for a given transition [22]. For example, in Q-learning, an estimate of the optimal action value function is updated at every state transition [26]. In most RL algorithms, approximations are usually presented in tabular form which may become problematic with high-dimensional data. The implementation of deep neural networks [27] to RL (deep reinforcement learning, DRL) does not require tabular representations for policies, value functions or Q. Recently, Mnih et al. [17] introduced deep Q-learning (DQL) that utilizes neural networks to directly learn policies from high-dimensional data and thereby overcomes the previous shortcomings of RL with neural nets by stabilizing the training of the action value function in an end-to-end RL approach while providing an algorithm that adapts to a variety of tasks (Atari games). More recently, Schrittwieser et al. [19] introduced MuZero that outperforms previous DRL algorithms in gameplay. In contrast to previous DRL algorithms, MuZero does not aim at modeling the entire environment but only models what is needed for the agent's decision-making value, policy and reward—using a deep neural network and tree-based search.

These recent advances, especially in DRL, that are adaptive to an increasing range of settings without the need to fully disclose the underlying dynamics of an environment, i.e., the rules of the game, provide a vast potential for applications in oncology where an abundance of high-dimensional data and rapid environmental changes limited previous efforts.

3. Recent Studies of Reinforcement Learning in Malignant Disease

Treatment regimens in oncology are usually longitudinal decision-making processes where patient variables as well as response to treatment and toxicities influence the oncologist's choice in order to optimize patient safety and outcome in the long run. This clinical framework can be translated into a set of sequential actions in an environment that result in iterative state alterations. In that sense, dynamic treatment regimens (DTRs) [28] can be set up as an RL problem due to its sequential nature, and dose adjustments can be performed by a digital agent that receives rewards for favorable events such as tumor response or curation and penalties for unfavorable events such as toxicities (Figure 3) [29]. Due to obvious ethical concerns in a trial-and-error learning method, RL in DTRs is usually applied, at present, in a retrospective setting or with simulated data based on historical cohorts.

Several recent studies have applied this approach for optimizing chemotherapy dosages, most commonly using Q-learning in simulated environments (Table 3). Padmanabhan et al. [30] employed Q-learning for chemotherapy dosing in an in silico approach in simulated patients in a closed loop to maximize on-target drug effects and minimize off-target toxicities. Additionally, utilizing Q-learning, Zade et al. [31] proposed a simulation framework where an RL agent optimizes the dosage of temozolomide in order to minimize glioblastoma tumor size. Yazdjerdi et al. [32] applied Q-learning to optimize anti-angiogenic therapy in a simulated tumor environment. RL-based drug sensitivity screening regarding different tumor cell lines with Q-rank was proposed by Daoud et al. [33]. Their method ranks drug sensitivity prediction algorithms and recommends the optimal algorithms for a given drug–cell line pair in order to achieve optimal responses. To account for chemotherapy-associated toxicity, Maier et al. [34] proposed an RL-based framework that is guided by absolute neutrophil counts for adjusting subsequent drug doses. Using simulated reinforcement trials [35], Zhao et al. [36] applied Q-learning to stage IIIB/IV non-small cell lung cancer and reported optimized first and second treatment lines as well as optimal selection for initiating second-line therapy. Similarly, Yauney et al. [37] aimed to minimize mean tumor diameters in a simulated trial of patients receiving chemo- and/or radiotherapy using action-derived rewards as approximations of patient outcome. Both Liu et al. [38] and Krakow et al. [39] used registry data from patients with hematologic malignancies who underwent allogeneic stem cell transplantation that were listed in the Center for International Blood and Marrow Transplant Research registry and applied DRL and Q-learning, respectively, in order to prevent and treat graft-versus-host disease.

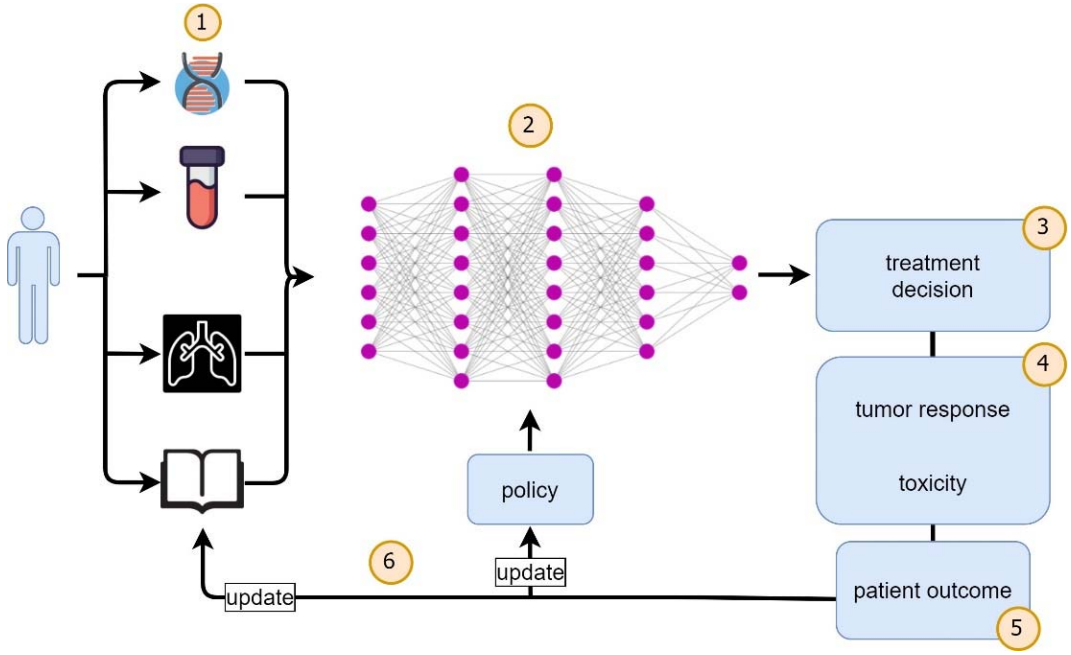

Figure 3. Iterative workflow of a reinforcement learning approach to precision oncology. For the individual patient, multimodal data (1), e.g., from genetic assays, laboratory tests, radiographic images and electronic health records, serve as an input for a reinforcement learning (RL) framework (2)—here, depicted as a deep neural net—which selects an action such as a treatment decision (3) according to its policy. This treatment decision will affect tumor response and toxicity (4) simultaneously and thus, ultimately, affect long-term patient outcome (5). This is translated into a reward signal for the RL agent which results in a policy update (6). At the same time, the state of the patient changes. For example, tumor response could be measured with radiographic imaging, and/or toxicity could be monitored by laboratory values and documentation in the electronic health record. This update of the state initiates a new cycle where updated inputs to the RL framework lead to a new treatment decision according to the updated policy, and the loop is closed.

Table 3. Recent studies of reinforcement learning (RL) for adaptive dosing of antineoplastic drugs in cancer.

Reference	Main Goal	Environment/ Cohort	Model-Based	Model-Free	V (State-Based)	Q (Action-Based)	Markov Assumption	No Markov Assumption	Table-/Map-Based	Deep Learning	Code Availability
[30]	Evaluation of an RL-based drug controller to enhance therapeutic effect on simulated tumors while sparing normal tissue without the necessity to disclose underlying system dynamics to the RL agent	15 simulated cancer patients		X		X	X		X		
[31]	Comparison of an RL-guided temozolomide treatment schedule to conventional clinical regimen	simulated glioblastoma tumor growth model	X			X	X		X		
[32]	RL-based optimization of anti-angiogenic therapy with endostatin in a simulated tumor growth model with dynamic patient parameters	simulated tumor growth model, simulated patient		X		X	X		X		X
[33]	Prediction of chemotherapy sensitivity in breast cancer cell lines with available multi-omics data by ranking suitable prediction algorithms using Q-rank	drug sensitivity data of 53 breast cancer cell lines	X			X	X		X		X

Table 3. *Cont.*

Reference	Main Goal	Environment/ Cohort	Model-Based	Model-Free	V (State-Based)	Q (Action-Based)	Markov Assumption	No Markov Assumption	Table-/Map-Based	Deep Learning	Code Availability
[34]	Evaluation of data assimilation techniques in combination with RL for dose adjustments of chemotherapy in simulated patients using absolute neutrophile count as a surrogate endpoint	simulated patients	X		X		X		X		X
[36]	RL-based dose adjustments for chemotherapy and initiation of second-line therapy while accounting for patient censoring	simulated clinical trial of stage IIIB/IV non-small cell lung cancer patients		X		X		X	X		
[37]	Deep RL-guided dosing regimens with temozolomide or procarbazine, CCNU and vincristine using action-derived rewards	simulated clinical trial using a glioblastoma tumor growth model	X			X	X		X		
[38]	Evaluation of RL-guided prevention and treatment of acute and chronic graft-versus-host disease	registry data from 6021 AML patients who underwent allogeneic stem cell transplantation		X		X		X		X	
[39]	Evaluation of RL-guided prevention and treatment of acute and chronic graft-versus-host disease	registry data from 11,141 patients who underwent allogeneic stem cell transplantation	X			X		X	X		

In parallel to chemotherapy regimens, RL can be applied to optimize radiotherapy to maximize on-target effects and minimize off-target toxicities (Table 4). Treatment planning and manual target segmentation still require an excessive amount of manual labor and time [40]. Deep learning has been investigated in order to aid the radiotherapist in treatment planning and reduce inter-observer variability. For example, different variations of convolutional neural nets have been developed for fast and accurate segmentation of brain metastases [41,42], thoracic cancer manifestations [43] or rectal cancer [44]. Accordingly, RL can be used for radiation dose adjustments for the individual patient. Kim et al. [45] defined the radiotherapeutic fractionation schedule as an MDP and proposed adaptive fractions according to individual patient response. Tseng et al. [46] used deep Q-learning to develop adaptive radiation protocols for patients with non-small cell lung cancer, balancing rewards for the agent between on-target efficiency and off-target toxicity. They accommodated for the initially small sample size with simulated patient data generated by a generative adversarial net (GAN). Jalalimanesh et al. [47] used an agent-based model and Q-learning to adapt fraction sizes to tumor response in a simulated environment. Similarly, adjustment of dose fractionation performed by a DRL agent in a simulated model of tumor growth was also demonstrated by Moreau et al. [48], who reported an improved performance compared to the baseline treatment plans. Using historic data from prostate cancer patients, Hrinivich et al. [49] applied deep Q-learning for volumetric modulated arc therapy and reported on-target and off-target doses comparable to clinical plans. Correspondingly, Shen et al. [50] used DRL in a virtual environment to generate treatment plans by training on 10 and validating on 64 cases of patients with prostate cancer. In a

similar approach, Zhang et al. [51] trained an RL agent on augmented treatment plans of 16 previously treated patients that received pancreas stereotactic body radiation therapy which was validated on 24 treatment plans, achieving a treatment quality comparable to clinical plans. It is to be noted that while the majority of the presented studies describe their algorithms in great mathematical detail, the information about the general problem setup and algorithm architecture has to be easily accessible to both software engineers and clinicians. In order to transparently report the used methodologies, authors of future studies of RL in medicine can refer to the proposed items in Table 2 for preparation of their paper's method section. This framework can help both the authors in clearly structuring their reports and the readers in effortlessly picking out the main components of a novel algorithm architecture for a given use case. Using such a standardized approach can help facilitate the reproducibility of RL research in medicine and may aid in transferring RL algorithms from one application in oncology to another.

Table 4. Recent studies of reinforcement learning (RL) for adaptive dosing and fractionation of radiotherapy in cancer.

Reference	Main Goal	Environment/ Cohort	Model-Based	Model-Free	V (State-Based)	Q (Action-Based)	Markov Assumption	No Markov Assumption	Table-/Map-Based	Deep Learning	Code Availability
[45]	Development of adaptive fractionation schemes based on mathematical modeling with a Markov decision process	Simulated environment of target volumes and organs at risk	X		X		X		X		
[46]	Evaluation of a multi-step deep learning model for radiation dose adjustments in a retrospective and augmented patient cohort compared to clinical treatment plans	Retrospective data of 114 non-small cell lung cancer patients and augmented data from a generative adversarial net		X		X		X		X	
[47]	Proof of concept of an RL agent for adaptive irradiation dosing and fractionation schemes	Simulated tumor growth model		X		X		X	X		
[48]	Comparison of adaptive dose fractionation schemes to clinical treatment regimens	Simulated tumor growth model	X			X	X			X	X
[49]	RL to guide volumetric modulated arc therapy with machine parameter optimization and comparison between on-target and off-target doses	Retrospective data of 40 patients with prostate cancer	X			X		X		X	
[50]	Training and evaluation of a RL-based deep virtual treatment planner	Retrospective data of 74 patients with prostate cancer		X		X		X		X	
[51]	Optimization of on-target and off-target dosing for stereotactic body irradiation in pancreatic cancer	Retrospective data of 16 patients with pancreatic cancer		X		X		X	X		

4. Discussion

The presented studies underline the feasibility of RL-guided precision oncology both regarding irradiation and drug therapy. However, the majority of previous studies suffer from common obstacles. In this section, we aim to highlight frequent challenges in RL

design for clinical use cases and discuss possible strategies to overcome these hurdles. Accurately mapping the environment and assessing the complexity of available data as well as the sequential nature of a clinical problem are the key first steps in setting up an RL support system. Biological systems and their behavior under environmental influences represent a highly complex system with a myriad of unknown variables in the context of disease and treatment. Hence, a detailed model of a clinical problem can often not be obtained. Several studies address this issue by using simplified simulations of tumor behavior [30–32,36,37,46,47]; however, the majority of these studies worked with relatively small samples which can limit the agent's capability of abstracting an efficient policy given few examples to train on [52]. This leads to the question of to what extent such algorithms are generalizable to more complex environments or real-life applications. Sparse and missing data are all too common in medical datasets. If the agent cannot access all information that is critical for decision making, a concluding model may misrepresent the actual environment. In that sense, most scenarios in clinical oncology behave in a non-Markovian way as not all relevant information is disclosed to the agent (or the clinician) [53]. Adding to the complexity, medical data may be biased or noisy due to inter-rater variability depending on the data source which may add to the variance of estimates of the value function and therefore affect policy determination [29]. While many cases of missing data in RL in general may be tackled with a partially observable Markov decision process design [54], the high dimensionality and complexity of medical data demand more sophisticated methods, such as multiple imputation models [55] or advanced Q-learning techniques for patients lost to follow-up [56]. Small sample sizes can be accounted for by pooling multicenter datasets which may, in turn, add variability to the dataset. Hence, standardization of data collection across institutions and even countries seems warranted to generate large high-quality datasets for future ML applications. To maintain high quality in such multicenter and multinational datasets, standardization of reporting as well as public access is essential. Internationally acclaimed frameworks for tumor response such as RECIST [57] or the reporting of adverse events such as CTCAE [58] as well as data from electronic health records [59] can be utilized to store clinical information in such datasets in a universally accessible way without the need for excessive pre-processing before pooling data from different sources. Data sharing between institutions and countries is crucial to create larger datasets, even for rare entities, and provide RL agents (and ML in general) with bigger sample sizes to train on. A frequent shortcoming of the studies cited above (with a few exceptions) is the lack of publicly available datasets and code. Often, only mathematical modeling or pseudocode is reported. However, to ensure reproducibility, public availability of both data and code is key. This will allow for independent model improvement, pooling of similar datasets and, overall, a faster pace and higher generalizability of RL models in oncology. Publishers should acknowledge this shortcoming and incentivize authors to share their code upon publication. However, informed patient consent about the processing of data and safety measures to protect patient identity need to be implemented. As this process naturally requires collaborative efforts and time, alternative approaches are needed for small data. GANs [60] can be implemented to augment small datasets. Their feasibility to add data to RL has recently been demonstrated in a dataset of patients with non-small cell lung cancer [46]. However, a study evaluating RL performance in a comparison between real-world and simulated data is lacking. Such a comparison could be made between a dataset generated by GANs and retrospective patient data in order to show discrepancies based on the data structure. This would allow for an in-depth look at the quality of simulated data which, in turn, could be improved to allow for a more robust simulation in future models. Another possibility is to first train the RL agent by expert demonstration, inverse learning, transfer learning or a combination thereof. Formulating a reward function a priori and then letting the agent derive an optimal policy may not be ubiquitously possible in a clinical setting with many unknown variables, and hence retrospective data of (near-)optimal treatment histories can be utilized to estimate a reward function based on previous expert decisions [29]. This can be achieved by

behavioral cloning, where pairs of environmental states and expert actions are mapped directly by the agent [61,62] (in a way similar to supervised learning), or by inverse RL, where a reward function is determined based on observing ideal decisions [63,64]. Still, it needs to be considered that in this scenario, the reward is based on a match between the agent's and the expert's decision, which may, again, result in bias as different experts may disagree over different decisions, and therefore misrepresentative rewards can result in poor performance and safety issues [65]. This leads to a fundamental issue at the heart of RL: credit assignment. The main incentive to reinforce an agent's behavior is encoded in the reward function, and henceforth, the reward signal determines whether or not a certain behavior is reinforced given a certain state of the environment. While this may be straightforward in gameplay where all underlying dynamics are known and the reward is often a direct consequence of an agent's action, rewards in a healthcare domain may be sparse, and the time between an action and its result may be considerably longer. In oncology, treatment effects evidently do not manifest themselves immediately, and linking an agent's action, e.g., a dose modification, to a certain outcome, e.g., prolonged relapse-free survival, remains challenging. Consequently, long-term rewards should be favored over short-term rewards, and oversimplifying reward functions can lead to unwanted behaviors, resulting in an agent doing more harm than good [52]. Furthermore, in comparison to gameplay where there usually is one single goal (win the game), oncologic practice demands a variety of treatment goals to be met such as improving survival, reducing morbidity, reducing toxicity and improving quality of life, among others. A possible way to deal with sparse rewards in the light of multiple goals is hindsight experience replay, where different learning episodes are replayed with different goals and the agent can derive reward signals regarding different outcomes [66]. In most applications of RL in healthcare, rewards are coded quantitatively rather than qualitatively, which can be useful for certain use cases where the outcome, in fact, is a metric variable (such as absolute neutrophile count [34]); however, it remains challenging when the outcome first has to be transformed or a priori model building has to be performed manually [29]. Alternatively, preference models can be used as a representation of qualitative feedback to rank the agent's behavioral trajectories [67,68]. However, a critical question is whether the reward an agent receives for an action is actually the optimal possible reward. This leads to another fundamental issue in RL, the trade-off between exploitation and exploration. Essentially, an agent has two options: either exploit current knowledge in order to achieve rewards or explore for previously unknown information which potentially leads to improved policies to gain higher rewards [22]. In the healthcare domain, especially in oncology, this dualism is crucial since exploration methods with insufficient safety measures can lead to potentially devastating outcomes, while insufficient exploration may lead to suboptimal policies and thus to unsatisfactory treatment decisions. Penalizing an agent for an unfavorable action may be insufficient when it comes to safety concerns in a healthcare setting, especially when the action leads to unrecoverable damage. This becomes especially relevant when dealing with drugs that have narrow therapeutic ranges and information on dose adaptation is limited [69]. Adding to the aforementioned challenge of multiple objectives in a healthcare setting is the fact that some objectives may be contradictory. For example, a full dose of chemotherapy may result in improved tumor response but, at the same time, will inevitably increase toxicity. A method to account for such contradictions is multi-objective RL that aims to evaluate polar objectives by obtaining a policy that represents Pareto optimal solutions [70]. While a lost game can simply be reset and started anew, an overdose in a clinical setting can potentially cost a patient's life. Hence, safe exploration strategies [71], especially in online learning, are crucial for RL in oncology. This raises the question of what the optimal benchmark should be when it comes to evaluating an RL agent's performance. Frequently, RL decisions are compared to clinical treatment plans; however, it remains questionable whether this is the optimal strategy since, conceivably, RL performance in a narrow domain could, at some point, exceed human performance in terms of decision making as it already has done, for example, in chess. When it comes to decision support

systems, safety goes hand in hand with trust. Let us assume that your RL algorithm suggests a dose alteration for a given patient. Do you trust that decision? If so, why? If not, why not? A major drawback of many current ML applications in such delicate environments as healthcare is interpretability, and DL in particular is often referred to as a 'black box' when it comes to exactly how an algorithm arrives at a conclusion [72,73]. This becomes especially challenging when the oncologic expert and the RL algorithm arrive at different solutions for the same problem [74,75]. Often, the signals an algorithm uses for decision making and the policies that are learnt can neither be accessed easily nor interpreted comprehensively by the human investigator [76]. Yet, the path to the conclusion is equally as important as the conclusion itself, especially in healthcare, where not only scientific knowledge gains are expected but patients also have an inherent right to be well-informed with respect to the background of a treatment decision. The interpretability of such RL algorithms should refrain from unnecessary abstraction and highlight causal pathways [77] that are meaningful to both the clinician and the patient. In that sense, understanding the exact model may be unnecessary in practice (to the clinician and patient) when causal pathways can be well interpreted. However, this is still an ongoing endeavor in ML in general [78,79], and satisfactory solutions tailored for healthcare applications are lacking [80], which bears the risk of reintroducing a paternalistic system in patient care [81]. Still, this remains controversial as it can be argued that the input from 'black box' systems is already happening to some extent in clinical oncology and is widely accepted in daily practice: hardly anyone seriously questions the results of molecular analysis or the assessment of biomarkers when it comes to clinical decision support, and confidence in these techniques has been built over recent years by reliable performance [82]. It is therefore conceivable that RL-based decision support systems, once they have been broadly tested and validated, may gain a similar level of trust as advanced biomedical techniques. In that regard, the frequent notion that ML systems could threaten the clinician's autonomy can be set aside as it is far more likely that these systems will be integrated as decision support in the same way that molecular and genetic data are implemented now, guiding precision oncology and further individualizing patient care, while the final responsibility for any taken decision undoubtedly lies with the oncologist. Previous studies focused on either conventional chemotherapy regimens or radiotherapy. However, the implementation of targeted therapy or immunotherapy in the treatment guidelines of many tumor entities calls for studies that account for these therapeutics as well and evaluate combinations of chemo-, radio- and targeted therapy in the respective tumor entities. These studies and algorithms have to be designed with diligence to both accurately map a clinically relevant problem setup in oncology and, at the same time, account for multiple different objectives and potential adverse effects in the context of multimodal contemporary therapy regimens.

5. Conclusions

RL in oncology is still in its infancy, and as we pointed out, a multitude of issues have to be properly addressed in future studies for these techniques to mature and find acceptance in clinical oncology. The sequential nature of RL and its capability for long-term outcome optimization make it a suitable candidate to be implemented in precision oncology, harnessing the growing body of available biomedical data for the individual patient. To progress in this potentially practice-changing field, an interdisciplinary effort to iteratively refine these systems for specific use cases as well as institutional guidelines is needed in order to achieve meaningful representations of clinically relevant tasks for optimal patient care.

Author Contributions: J.-N.E.: conceptualization, literature search, visualization, writing—original draft, validation. K.W.: writing—review and editing, validation. M.B.: writing—review and editing, validation. J.M.M.: funding acquisition, project administration, supervision, writing—review and editing, validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Technical University Dresden, grant number 60499 to Middeke, and a scholarship by the Deutsche Krebshilfe (Mildred-Scheel Nachwuchszentrum, Dresden, Germany) to Eckardt. The Else-Kroener Fresenius Centre for Digital Health (EKFZ) is acknowledged for supporting the AI initiative at TU Dresden.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef]
2. He, J.; Baxter, S.L.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nat. Med.* **2019**, *25*, 30–36. [CrossRef]
3. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson: Boston, MA, USA, 2020; ISBN 978-0-13-461099-3.
4. Rodríguez-Ruiz, A.; Krupinski, E.; Mordang, J.-J.; Schilling, K.; Heywang-Köbrunner, S.H.; Sechopoulos, I.; Mann, R.M. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* **2019**, *290*, 305–314. [CrossRef]
5. Perek, S.; Kiryati, N.; Zimmerman-Moreno, G.; Sklair-Levy, M.; Konen, E.; Mayer, A. Classification of Contrast-Enhanced Spectral Mammography (CESM) Images. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 249–257. [CrossRef]
6. Massafra, R.; Bove, S.; Lorusso, V.; Biafora, A.; Comes, M.C.; Didonna, V.; Diotaiuti, S.; Fanizzi, A.; Nardone, A.; Nolasco, A.; et al. Radiomic Feature Reduction Approach to Predict Breast Cancer by Contrast-Enhanced Spectral Mammography Images. *Diagnostics* **2021**, *11*, 684. [CrossRef] [PubMed]
7. Amoroso, N.; Pomarico, D.; Fanizzi, A.; Didonna, V.; Giotta, F.; La Forgia, D.; Latorre, A.; Monaco, A.; Pantaleo, E.; Petruzzellis, N.; et al. A Roadmap towards Breast Cancer Therapies Supported by Explainable Artificial Intelligence. *Appl. Sci.* **2021**, *11*, 4881. [CrossRef]
8. Comes, M.C.; La Forgia, D.; Didonna, V.; Fanizzi, A.; Giotta, F.; Latorre, A.; Martinelli, E.; Mencattini, A.; Paradiso, A.V.; Tamborra, P.; et al. Early Prediction of Breast Cancer Recurrence for Patients Treated with Neoadjuvant Chemotherapy: A Transfer Learning Approach on DCE-MRIs. *Cancers* **2021**, *13*, 2298. [CrossRef]
9. Dembrower, K.; Wåhlin, E.; Liu, Y.; Salim, M.; Smith, K.; Lindholm, P.; Eklund, M.; Strand, F. Effect of Artificial Intelligence-Based Triaging of Breast Cancer Screening Mammograms on Cancer Detection and Radiologist Workload: A Retrospective Simulation Study. *Lancet Digit. Health* **2020**, *2*, e468–e474. [CrossRef]
10. Jafari, M.; Wang, Y.; Amiryousefi, A.; Tang, J. Unsupervised Learning and Multipartite Network Models: A Promising Approach for Understanding Traditional Medicine. *Front. Pharmacol.* **2020**, *11*, 1319. [CrossRef] [PubMed]
11. Awada, H.; Durmaz, A.; Gurnari, C.; Kishtagari, A.; Meggendorfer, M.; Kerr, C.M.; Kuzmanovic, T.; Durrani, J.; Shreve, J.; Nagata, Y.; et al. Machine Learning Integrates Genomic Signatures for Subclassification Beyond Primary and Secondary Acute Myeloid Leukemia. *Blood* **2021**. [CrossRef]
12. Yu, K.-H.; Beam, A.L.; Kohane, I.S. Artificial Intelligence in Healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef]
13. Campbell, M.; Hoane, A.J.; Hsu, F. Deep Blue. *Artif. Intell.* **2002**, *134*, 57–83. [CrossRef]
14. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484–489. [CrossRef]
15. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play. *Science* **2018**, *362*, 1140–1144. [CrossRef] [PubMed]
16. Bellemare, M.G.; Naddaf, Y.; Veness, J.; Bowling, M. The Arcade Learning Environment: An Evaluation Platform for General Agents. *J. Artif. Intell. Res.* **2013**, *47*, 253–279. [CrossRef]
17. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
18. Vinyals, O.; Babuschkin, I.; Czarnecki, W.M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D.H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning. *Nature* **2019**, *575*, 350–354. [CrossRef]
19. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature* **2020**, *588*, 604–609. [CrossRef] [PubMed]
20. Sallab, A.E.; Abdou, M.; Perot, E.; Yogamani, S. Deep Reinforcement Learning Framework for Autonomous Driving. *Electron. Imaging* **2017**, *2017*, 70–76. [CrossRef]
21. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*; The National Academies Collection: Reports funded by National Institutes of Health; National Academies Press (US): Washington, DC, USA, 2011; ISBN 978-0-309-22222-8.
22. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018; ISBN 978-0-262-35270-3.
23. Li, Y. Deep Reinforcement Learning: An Overview. *arXiv* **2018**, arXiv:1701.07274.

24. Travassos, G.; Barros, M. Contributions of In Virtuo and In Silico Experiments for the Future of Empirical Studies in Software Engineering Contributions of In Virtuo and In Silico Experiments for the Future of Empirical Studies in Software Engineering. In *The Future of Empirical Studies in Software Engineering: Proceedings of the ESEIW 2003 Workshop on Empirical Stuides in Software Engineering (WSESE 2003), Rome, Italy, 29 September 2003*; Springer: Berlin/Heidelberg, Germany, 2004; Volume 2, p. 117.
25. Jonsson, A. Deep Reinforcement Learning in Medicine. *KDD* **2019**, *5*, 18–22. [CrossRef] [PubMed]
26. Watkins, C.; Dayan, P. Q-Learning. *Proc. Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
27. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
28. Chakraborty, B.; Murphy, S.A. Dynamic Treatment Regimes. *Annu. Rev. Stat. Its Appl.* **2014**, *1*, 447–464. [CrossRef]
29. Yu, C.; Liu, J.; Nemati, S. Reinforcement Learning in Healthcare: A Survey. *arXiv* **2020**, arXiv:1908.08796.
30. Padmanabhan, R.; Meskin, N.; Haddad, W.M. Reinforcement Learning-Based Control of Drug Dosing for Cancer Chemotherapy Treatment. *Math. Biosci.* **2017**, *293*, 11–20. [CrossRef]
31. Ebrahimi Zade, A.; Shahabi Haghighi, S.; Soltani, M. Reinforcement Learning for Optimal Scheduling of Glioblastoma Treatment with Temozolomide. *Comput. Methods Programs Biomed.* **2020**, *193*, 105443. [CrossRef]
32. Yazdjerdi, P.; Meskin, N.; Al-Naemi, M.; Al Moustafa, A.-E.; Kovács, L. Reinforcement Learning-Based Control of Tumor Growth under Anti-Angiogenic Therapy. *Comput. Methods Programs Biomed.* **2019**, *173*, 15–26. [CrossRef]
33. Daoud, S.; Mdhaffar, A.; Jmaiel, M.; Freisleben, B. Q-Rank: Reinforcement Learning for Recommending Algorithms to Predict Drug Sensitivity to Cancer Therapy. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3154–3161. [CrossRef]
34. Maier, C.; Hartung, N.; Kloft, C.; Huisinga, W.; Wiljes, J. de Reinforcement Learning and Bayesian Data Assimilation for Model-Informed Precision Dosing in Oncology. *CPT Pharmacomet. Syst. Pharmacol.* **2021**, *10*, 241–254. [CrossRef] [PubMed]
35. Zhao, Y.; Kosorok, M.R.; Zeng, D. Reinforcement Learning Design for Cancer Clinical Trials. *Stat. Med.* **2009**, *28*, 3294–3315. [CrossRef] [PubMed]
36. Zhao, Y.; Zeng, D.; Socinski, M.A.; Kosorok, M.R. Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer. *Biometrics* **2011**, *67*, 1422–1433. [CrossRef]
37. Yauney, G.; Shah, P. Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection. In Proceedings of the Machine Learning for Healthcare Conference, PMLR, Stanford, CA, USA, 16–18 August 2018; pp. 161–226.
38. Liu, Y.; Logan, B.; Liu, N.; Xu, Z.; Tang, J.; Wang, Y. Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data. *Healthc. Inform.* **2017**, *2017*, 380–385. [CrossRef] [PubMed]
39. Krakow, E.F.; Hemmer, M.; Wang, T.; Logan, B.; Arora, M.; Spellman, S.; Couriel, D.; Alousi, A.; Pidala, J.; Last, M.; et al. Tools for the Precision Medicine Era: How to Develop Highly Personalized Treatment Recommendations From Cohort and Registry Data Using Q-Learning. *Am. J. Epidemiol.* **2017**, *186*, 160–172. [CrossRef]
40. Boldrini, L.; Bibault, J.-E.; Masciocchi, C.; Shen, Y.; Bittner, M.-I. Deep Learning: A Review for the Radiation Oncologist. *Front. Oncol.* **2019**, *9*, 977. [CrossRef]
41. Liu, Y.; Stojadinovic, S.; Hrycushko, B.; Wardak, Z.; Lau, S.; Lu, W.; Yan, Y.; Jiang, S.B.; Zhen, X.; Timmerman, R.; et al. A Deep Convolutional Neural Network-Based Automatic Delineation Strategy for Multiple Brain Metastases Stereotactic Radiosurgery. *PLoS ONE* **2017**, *12*, e0185844. [CrossRef]
42. Charron, O.; Lallement, A.; Jarnet, D.; Noblet, V.; Clavier, J.-B.; Meyer, P. Automatic Detection and Segmentation of Brain Metastases on Multimodal MR Images with a Deep Convolutional Neural Network. *Comput. Biol. Med.* **2018**, *95*, 43–54. [CrossRef]
43. Trullo, R.; Petitjean, C.; Ruan, S.; Dubray, B.; Nie, D.; Shen, D. Segmentation of Organs at Risk in Thoracic CT Images Using a Sharpmask Architecture and Conditional Random Fields. *Proc. IEEE Int. Symp. Biomed. Imaging* **2017**, *2017*, 1003–1006. [CrossRef] [PubMed]
44. Men, K.; Dai, J.; Li, Y. Automatic Segmentation of the Clinical Target Volume and Organs at Risk in the Planning CT for Rectal Cancer Using Deep Dilated Convolutional Neural Networks. *Med. Phys.* **2017**, *44*, 6377–6389. [CrossRef] [PubMed]
45. Kim, M.; Ghate, A.; Phillips, M.H. A Markov Decision Process Approach to Temporal Modulation of Dose Fractions in Radiation Therapy Planning. *Phys. Med. Biol.* **2009**, *54*, 4455–4476. [CrossRef]
46. Tseng, H.-H.; Luo, Y.; Cui, S.; Chien, J.-T.; Haken, R.K.T.; Naqa, I.E. Deep Reinforcement Learning for Automated Radiation Adaptation in Lung Cancer. *Med. Phys.* **2017**, *44*, 6690–6705. [CrossRef]
47. Jalalimanesh, A.; Shahabi Haghighi, H.; Ahmadi, A.; Soltani, M. Simulation-Based Optimization of Radiotherapy: Agent-Based Modeling and Reinforcement Learning. *Math. Comput. Simul.* **2017**, *133*, 235–248. [CrossRef]
48. Moreau, G.; François-Lavet, V.; Desbordes, P.; Macq, B. Reinforcement Learning for Radiotherapy Dose Fractioning Automation. *Biomedicines* **2021**, *9*, 214. [CrossRef]
49. Hrinivich, W.T.; Lee, J. Artificial Intelligence-Based Radiotherapy Machine Parameter Optimization Using Reinforcement Learning. *Med. Phys.* **2020**, *47*, 6140–6150. [CrossRef]
50. Shen, C.; Nguyen, D.; Chen, L.; Gonzalez, Y.; McBeth, R.; Qin, N.; Jiang, S.B.; Jia, X. Operating a Treatment Planning System Using a Deep-Reinforcement Learning-Based Virtual Treatment Planner for Prostate Cancer Intensity-Modulated Radiation Therapy Treatment Planning. *Med. Phys.* **2020**, *47*, 2329–2336. [CrossRef] [PubMed]

51. Zhang, J.; Wang, C.; Sheng, Y.; Palta, M.; Czito, B.; Willett, C.; Zhang, J.; Jensen, P.J.; Yin, F.-F.; Wu, Q.; et al. An Interpretable Planning Bot for Pancreas Stereotactic Body Radiation Therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2021**, *109*, 1076–1085. [CrossRef] [PubMed]
52. Gottesman, O.; Johansson, F.; Komorowski, M.; Faisal, A.; Sontag, D.; Doshi-Velez, F.; Celi, L.A. Guidelines for Reinforcement Learning in Healthcare. *Nat. Med.* **2019**, *25*, 16–18. [CrossRef]
53. Coronato, A.; Naeem, M.; De Pietro, G.; Paragliola, G. Reinforcement Learning for Intelligent Healthcare Applications: A Survey. *Artif. Intell. Med.* **2020**, *109*, 101964. [CrossRef]
54. Sondik, E.J. The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs. *Oper. Res.* **1978**, *26*, 282–304. [CrossRef]
55. Shortreed, S.M.; Laber, E.; Lizotte, D.J.; Stroup, T.S.; Pineau, J.; Murphy, S.A. Informing Sequential Clinical Decision-Making through Reinforcement Learning: An Empirical Study. *Mach. Learn.* **2011**, *84*, 109–136. [CrossRef]
56. Goldberg, Y.; Kosorok, M.R. Q-Learning with Censored Data. *Annu. Stat.* **2012**, *40*, 529–560. [CrossRef] [PubMed]
57. Schwartz, L.H.; Seymour, L.; Litière, S.; Ford, R.; Gwyther, S.; Mandrekar, S.; Shankar, L.; Bogaerts, J.; Chen, A.; Dancey, J.; et al. RECIST 1.1—Standardisation and Disease-Specific Adaptations: Perspectives from the RECIST Working Group. *Eur. J. Cancer* **2016**, *62*, 138–145. [CrossRef]
58. Trotti, A.; Colevas, A.D.; Setser, A.; Rusch, V.; Jaques, D.; Budach, V.; Langer, C.; Murphy, B.; Cumberlin, R.; Coleman, C.N.; et al. CTCAE v3.0: Development of a Comprehensive Grading System for the Adverse Effects of Cancer Treatment. *Semin. Radiat. Oncol.* **2003**, *13*, 176–181. [CrossRef]
59. Ross, M.K.; Wei, W.; Ohno-Machado, L. "Big Data" and the Electronic Health Record. *Yearb. Med. Inform.* **2014**, *23*, 97–104. [CrossRef] [PubMed]
60. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661.
61. Torabi, F.; Warnell, G.; Stone, P. Behavioral Cloning from Observation. *arXiv* **2018**, arXiv:1805.01954.
62. Ho, J.; Gupta, J.K.; Ermon, S. Model-Free Imitation Learning with Policy Optimization. *arXiv* **2016**, arXiv:1605.08478.
63. Ng, A.Y.; Russell, S. Algorithms for Inverse Reinforcement Learning. In Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000; Morgan Kaufmann: San Francisco, CA, USA, 2000; pp. 663–670.
64. Abbeel, P.; Ng, A.Y. Apprenticeship Learning via Inverse Reinforcement Learning. In Proceedings of the 21st International Conference on Machine learning, Banff, AB, Canada, 4–8 July 2004; Association for Computing Machinery: New York, NY, USA, 2004; p. 1.
65. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565.
66. Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Abbeel, P.; Zaremba, W. Hindsight Experience Replay. *arXiv* **2018**, arXiv:1707.01495.
67. Fürnkranz, J.; Hüllermeier, E. (Eds.) *Preference Learning*; Springer: Berlin/Heidelberg, Germany, 2011; ISBN 978-3-642-14124-9.
68. Wirth, C.; Fürnkranz, J.; Neumann, G. Model-Free Preference-Based Reinforcement Learning. *AAAI* **2016**, *30*, 2222–2228.
69. de Jonge, M.E.; Huitema, A.D.R.; Schellens, J.H.M.; Rodenhuis, S.; Beijnen, J.H. Individualised Cancer Chemotherapy: Strategies and Performance of Prospective Studies on Therapeutic Drug Monitoring with Dose Adaptation: A Review. *Clin. Pharmacol.* **2005**, *44*, 147–173. [CrossRef]
70. Liu, C.; Xu, X.; Hu, D. Multiobjective Reinforcement Learning: A Comprehensive Overview. *IEEE Trans. Syst. Man Cybern. Syst.* **2015**, *45*, 385–398. [CrossRef]
71. García, J.; Fernández, F. A Comprehensive Survey on Safe Reinforcement Learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
72. Castelvecchi, D. Can We Open the Black Box of AI? *Nat. News* **2016**, *538*, 20. [CrossRef] [PubMed]
73. Lipton, Z.C. The Mythos of Model Interpretability. *arXiv* **2017**, arXiv:1606.03490.
74. Grote, T.; Berens, P. On the Ethics of Algorithmic Decision-Making in Healthcare. *J. Med. Ethics* **2020**, *46*, 205–211. [CrossRef]
75. Azuaje, F. Artificial Intelligence for Precision Oncology: Beyond Patient Stratification. *NPJ Precis. Oncol.* **2019**, *3*, 1–5. [CrossRef]
76. Humphreys, P. The Philosophical Novelty of Computer Simulation Methods. *Synthese* **2009**, *169*, 615–626. [CrossRef]
77. Madumal, P.; Miller, T.; Sonenberg, L.; Vetere, F. Explainable Reinforcement Learning through a Causal Lens. *AAAI* **2020**, *34*, 2493–2500. [CrossRef]
78. Molnar, C. *Interpretable Machine Learning*; Lulu: Morrisville, NC, USA, 2020; ISBN 978-0-244-76852-2.
79. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
80. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable Artificial Intelligence: A Survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 210–215.
81. McDougall, R.J. Computer Knows Best? The Need for Value-Flexibility in Medical AI. *J. Med. Ethics* **2019**, *45*, 156–160. [CrossRef] [PubMed]
82. Nardini, C. Machine Learning in Oncology: A Review. *Ecancermedicalscience* **2020**, *14*, 1065. [CrossRef] [PubMed]