Bias in Reinforcement Learning: A Review in Healthcare Applications

BENJAMIN SMITH, University of Tennessee: Bredesen Center for Interdisciplinary Research
ANAHITA KHOJANDI, University of Tennessee: Department of Industrial and Systems Engineering
RAMA VASUDEVAN, Oak Ridge National Laboratory: Center for Nanophase Materials Sciences

Reinforcement learning (RL) can assist in medical decision making using patient data collected in electronic health record (EHR) systems. RL, a type of machine learning, can use these data to develop treatment policies. However, RL models are typically trained using imperfect retrospective EHR data. Therefore, if care is not taken in training, RL policies can propagate existing bias in healthcare. Literature that considers and addresses the issues of bias and fairness in sequential decision making are reviewed. The major themes to mitigate bias that emerge relate to (1) data management; (2) algorithmic design; and (3) clinical understanding of the resulting policies.

CCS Concepts: • **Applied computing** → **Health care information systems**; **Health informatics;**

Additional Key Words and Phrases: Reinforcement learning, electronic health records, algorithmic bias, treatment planning, bias management

ACM Reference format:
Benjamin Smith, Anahita Khojandi, and Rama Vasudevan. 2023. Bias in Reinforcement Learning: A Review in Healthcare Applications. *ACM Comput. Surv.* 56, 2, Article 52 (September 2023), 17 pages.
https://doi.org/10.1145/3609502

1 INTRODUCTION

Reinforcement learning (RL) is a branch of **machine learning (ML)** focused on sequential decision making over time. In RL, an "agent" interacts with an "environment", and takes "actions" that can affect the "state" of the environment, and in turn, receives "rewards." The goal of RL is to learn an action strategy or "policy," i.e., a mapping of states to actions, that maximizes the total reward [56]. RL has traditionally been used in robotics and in game play, where the agent tries to navigate through or control actions within a well-defined game environment such as Pong or Space Invaders [52, 65]. Here, the agent attempts to learns best action at each time point that leads to winning the game or conversely, maximizing the reward (total points). While games have served as a foundation for RL research, it is now increasingly being applied to real-world problems that are focused on sequential decision making, given the growth of data volumes that make learning feasible.

The research is partially supported by Science Alliance, The University of Tennessee, and the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy.
Authors' addresses: B. Smith, Ph.D. and Graduate Researcher, University of Tennessee, Bredesen Center for Interdisciplinary Research, Knoxville, TN; A. Khojandi, Ph.D. Associate Professor, University of Tennessee, Department of Industrial and Systems Engineering, Knoxville, TN; R. Vasudevan, Ph.D. Group Leader, Oak Ridge National Laboratory, Center for Nanophase Materials Sciences, Oak Ridge, TN.

0360-0300/2023/09-ART52 $15.00
https://doi.org/10.1145/3609502

Recently, RL has shown great promise in healthcare applications, and in particular, in medical decision making. In healthcare, the goal for clinicians is generally to identify a treatment policy that maximizes the patients' benefit, e.g., **total expected quality adjusted lifetime (QALY)**. In the past, developing such treatment policies mostly relied on physician intuition and simple rule-based methods developed through clinical trials [8]. However, through the big data revolution, in many cases, researchers now have access to temporal measurements of different patient metrics over time that can be used to learn treatment policies using data-driven approaches. This is mostly enabled through the digital records of patient health information that are collected at the bedside in hospitals, especially in the data-rich **intensive care units (ICUs)**, and stored in **electronic health record (EHR)** systems. The data compiled in these systems are either entered by practitioners or other personnel, or are directly pulled into the EHRs using other means such as bedside monitors. The corresponding data track patient health status, the interventions made by practitioners, and the consequent patient responses to these interventions. Together, this information can create useful datasets of retrospective temporal data to train RL models [50].

Many RL studies exist that use EHR and other health data on an array of diseases, including AIDS [72], sepsis management [62, 82], Parkinson's disease [6], implanted cardiac device follow-up care [42, 43], and Heparin dosing [4]. In these studies, generally, RL agents leverage historical data to devise treatment policies that can be generalized to new patients. While the RL-prescribed policies are often not expected to be immediately and directly used in clinical settings, in the future they can provide new insights for practitioners and augment factors considered in medical decision making.

Although using RL in healthcare has the potential to provide new insights and improve patient outcomes, care should be taken to avoid *algorithmic bias*. Algorithms that make decisions differently among patient groups may exhibit algorithmic bias [1]. Recent research highlights the potential for algorithmic bias in supervised classification [57]. Some prominent examples include misidentifying faces of individuals from a certain race [49] or Warfarin dosing classification based on patient ethnicity [88, 92]. For ML-based classification problems, instances of algorithmic bias can often be readily observed due to the non-temporal nature of these problems [75]. As a branch of ML, RL problems are not immune to algorithmic bias, and yet, the issue of bias in RL has received considerably less attention in the literature. Here, bias is less obvious, but equally concerning as it can impact the course of treatment for patients. Hence, there is a timely need to investigate bias in RL and discuss the potential methods to identify and mitigate bias in practice.

2 BACKGROUND

Here, we discuss the fundamentals of RL and training RL with EHR data. We also describe applications of RL in healthcare and discuss active challenges.

2.1 RL Fundamentals

ML, an **artificial intelligence (AI)** method, can be implemented to detect patterns in data. This is largely possible through leveraging different function approximation methods [20]. Within ML exist different techniques and of these methods, RL applies to sequential decision making [56].

The main components of the RL system are an agent (or agents) and an environment. The agent(s) acts in the environment and learns the relationship between states and actions to maximize some rewards. In healthcare, the state of the environment can capture patient health state, e.g., blood pressure or the state of all vitals, and the action is the intervention performed, e.g., the amount and timing of **intravenous (IV)** medicine administration. Consequently, the rewards relate to the health outcomes of the patient, e.g., a successful discharge or normal vital sign measurements.

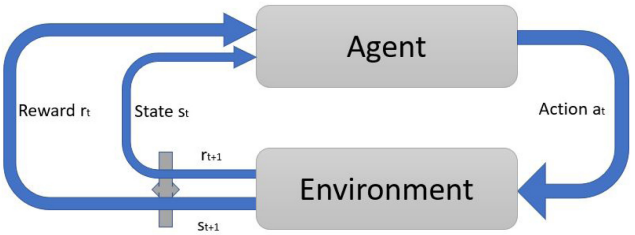

Fig. 1. An overview of the RL framework. An agent in state s_t takes action a_t. The environment then transitions to state s_{t+1} and receives the reward r_{t+1}. The process then continues starting from the new state.

To understand how an agent learns an environment and maximizes rewards, it is useful to understand the formal mathematical problem. **Markov decision processes (MDPs)** formalize the sequential agent and environmental interactions [70, 90]. A fully observable MDP is defined by a tuple of states, actions, transition function, reward function, and discount factor, $< S, A, P, R, \gamma >$. States, $s_t \in S$, and actions, $a_t \in A$, are used to create a transition probability function from the current state s_t to next state s_{t+1} and receive reward r after taking action a_t. This is given by $P(s_{t+1}, r]s_t, a_t)$ a conditional probability for each state and action pair [59]. The agent receives a reward (or incurs a cost) as a result of taking an action in a given state. The reward function $R(s_t, a_t) : S \times A \mapsto \mathbb{R}$ returns the immediate reward for each state and action pair [107]. By exploring states and actions, the agent attempts to find a policy to take an action in each state that maximizes the reward function. Specifically, the agent seeks to act to maximize the expected sum of discounted rewards where the discount factor, $\gamma \in [0, 1]$, is used to determine the tradeoff between valuing the immediate reward of taking an action over future expected rewards under the current policy. Figure 1 shows the overview of the RL framework.

RL models can be applied and solved in different ways. An RL algorithm that can use its own policy to generate samples is considered "on-policy." In on-policy learning, the agent interacts with the environment according to policy, π, that it attempts to improve. Otherwise, algorithms that learn from actions of previous policies are considered "off-policy" as they use other policies to develop a new one. Similarly, "off-line" RL occurs with no additional data collection, and when implemented, it solely uses historical/retrospective data. "Off-line" RL works with a fixed dataset, while "on-line" learning is able to learn from data as they are collected. The distributional shift between the historical data and the real environment presents a major challenge in "off-line" RL and oftentimes results in poor estimations of rewards for RL policies. However, some recent progress has been made [52].

2.2 Electronic Health Records (EHRs)

EHRs digitally store clinical data regarding patient health and can be leveraged by RL to evaluate past treatments and prescribe future treatment interventions for new patients [55, 87]. EHRs generally provide rich data, collected across different patients with a wide range of characteristics and co-morbidities. In addition, they typically include longitudinal data, which can enable the prediction of patient health state evolution under various interventions [27, 103].

EHRs usually contain high-dimensional time series data and static patient information with heterogeneous data types [76]. These datasets can describe patient demographic information and known pre-existing conditions across patients. Dynamic state info such as vital sign measurements, diagnostic codes, physician observations, and the potential treatment information are also often recorded [3, 61, 84]. While some collection is automatic and includes objective data

(e.g., data collected directly from heart rate monitors), other information is dependent on medical personnel. Manual data collection is prone to human factor errors [12, 85].

Various EHR datasets are currently publicly available, and have been successfully used in the past to develop RL models. Some of these datasets focus on the data-rich environments of ICU, e.g., MIMIC [39] and the eICU dataset [76]. The ICU is a particularly suitable setting for collecting high-frequency data and training and deploying RL models, due to continuous monitoring of patients (unless patients are moved for testing or imaging, etc.). In particular, these datasets have been heavily used in off-line RL research for improving treatment planning; they allow for examining past treatment actions and patient state information to evaluate current practices and explore potential improvements.

While EHRs can enable RL modeling, there are often caveats. Often the time increments for recording patient information is not consistent. Thus, data often contains many missing values, which makes working with the data more challenging and may require imputation of missing values [7]. Also, certain identifiers of patients are not always included because of ethical or privacy concerns [76]. EHRs also are affected by external factors not demonstrable in the data including logistical constraints for the hospital and physician availability for any given patient. In this way, external bias is introduced to EHRs data. There are efforts towards standardization of EHRs that include of social determinants of health, but this is not yet widely available [81]. Due to these issues, learning treatment policies from EHRs should consider how the data can differ from those of the actual patient populations, and care must be taken not to introduce bias into the RL-prescribed treatment policies.

2.3 RL in Healthcare

When using RL in healthcare applications, the state of the environment is generally set up to capture the patient's health state, provided through EHRs. The most comprehensive states provide the best description of the patient, and can prevent mistaken correlations of variables. Actions are then whether or not some treatment is provided at a certain level at a specific time. Transition probabilities are learned based on the observed transitions in the data from certain states to another, given the actions. Given data availability, states, and actions can be defined in a very granular manner. However, if needed, patient states can also be clustered to approximate policies for patients in similar medical states [97]. In addition, while discrete actions are more widely used in practice, continuous actions are be more representative of certain medical treatments [82]. Lastly, the agent attempts to learn interventions in each state to maximize rewards to improve the long-term health of a patient. A reward could be for maintaining some vital measurement within a certain range or for the patient getting discharged from the ICU to the floor, among others.

In healthcare, generally the goal is to develop a *clinician-in-the-loop* framework where RL acts as a decision support system [61]. That is, RL is trained to prescribe actions, given the observed state of the patient in real-time. However, treatment decisions are ultimately left to clinicians at the bedside rather than being automatically acted upon by the agent [93]. Figure 2 provides a schematic of this approach.

There have been numerous successful studies, adapting different RL techniques for healthcare applications [60]. This includes developing RL models that could be implemented in the ICU to improve glucose control of septic patients [78, 100], and to recommend the dosing of heparin, an anticoagulant medication [4, 26, 67]. Similarly, in the outpatient setting, RL has been paired with wearable sensor data that track motor functions in response to taking medications to optimize medication regimen for Parkinson's disease patients [6, 98].

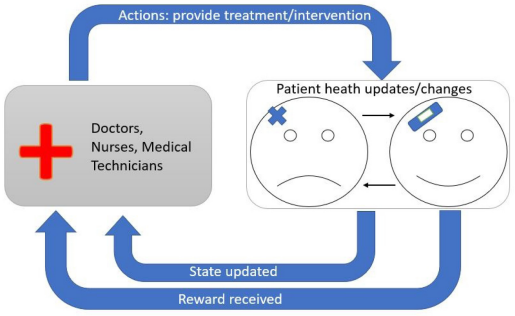

Fig. 2. RL in a healthcare setting. Similar to the traditional RL diagram, a clinician-in-the-loop RL framework is shown. Clinicians take an action in the form of treatment or testing that causes the patients state to transition. The patient's health could then change. Accordingly, the updated state and reward is returned to clinicians who repeat the process and take the appropriate actions at each time step.

2.4 Challenges

RL can play an important role in creating patient-specific treatment models. However, learning such data-driven models come with certain unique challenges. In creating a treatment policy, an RL agent needs sufficient data to describe patient states to learn patient-specific transitions. But healthcare data are often retrospective and clearly, running live experiments where the RL agent directly interacts with patients to collect more data is not ethically possible. Therefore, RL treatment models are typically trained in an off-line setting using existing EHRs.

As discussed in Section 2.2, the data used to train the model might not truly represent the patient population being treated for said disease. This directly relates to the "domain shift" problem in off-line RL; a potential patient population shift would decrease effectiveness of RL treatment policies due to overestimation of performance [91].

RL also deserves attention because it can make decisions based on more factors than considerable by a clinician [55]; however, this makes some decisions less understandable with human intuition. Explaining why RL recommends certain policies is difficult. Therefore, understanding whether the resulting patient-specific policies are accurate and equitable when put to use is extremely challenging [41, 77]. As of 2018, the EU protects individuals' rights to explanation in automated decision making [28]. RL policies should be explainable to people with diverse backgrounds [33] before they are effectively ready for adoption in practice.

Furthermore, it is often challenging to quantify how a certain patient is doing at any given point in time. RL transitions sometimes fail to capture a patient's entire health trajectory, which is further complicated due to patient covariates that may not be tracked in the EHR [93]. In addition, when RL rewards are based on medical measurements that are taken infrequently, this directly increases the difficulty in designing appropriate reward functions to guide the agent's learning. Because of these challenges, RL treatment policies must be carefully trained with the involvement of domain experts so as not to overstate algorithmic assumptions or impede the role of physicians.

3 LITERATURE REVIEW

In this section, we review the literature to draw insights as to the approaches that can help manage bias in healthcare.

3.1 Current Approaches to Bias Management in Healthcare

The medical field has considered differences in patient responses to treatments for years, directly or indirectly [2]. Early statistical studies developed policies that were optimized for different

patient sub-populations to demonstrate these differences [94]. Despite this, current medical practices may exhibit bias in treatments due to the heterogeneity of patients and their responses [51]. The potential for discrimination certainly poses challenges for the adoption of any new treatment plan in healthcare [71]. Healthcare bias is seen in unequal asset allocation or directly in treatment plan efficacy between patients [21].

Currently, there exist disparities in quality of care across patients, particularly as they relate to characteristics such as gender, race, ethnicity, and so on. [83]. For instance, in Veteran Affairs hospitals, black veterans are less likely to receive diagnostic testing for cardiovascular issues or undergo bypass procedures than their white counterparts [101]. This is just one example of how similar patients receive different treatments, even when financial costs are irrelevant.

In addition to differences in quality of care due to overt patient characteristics, non-clinical influences also affect medical decision-making [30]. For example, socio-economic status is directly correlated with the likelihood of receiving the best, and often most expensive, care [9]. Hospital equipment availability, physician time constraints, or the cost of treatment have shown to contribute to treatment decisions, adversely impacting poorer patients [53]. This does not mean that physicians are actively discriminating against patients but that external factors affect how clinical decisions are made [30]. The impact of these non-clinical influences on treatment actions and patient responses are further confounded by patient characteristics such as race, gender, or age [9, 10].

Classically, medical policies are developed through clinical trials [92]. In this setting, one way to reduce biased medical treatments is through sub-group analysis [45, 61]. Sub-groups may be used to develop different treatment models and also in analysis of treatment policies. Using sub-groups to develop different treatment plans can improve statistical effectiveness among patient populations using factors such as patient demographics [17]. Clinical trials use adaptive treatment strategies for individuals as an advanced way to differentiate patient groups [66]. Sequential multiple-assignment randomized trials incorporate initial patient responses to treatments to further define sub-groups to create different treatment policies.

Although sub-group analysis has shown the potential to reduce bias in treatment planning [2], such analysis is sometimes considered after the development process or is complicated due to the interaction between many factors. Despite statistical observations, sometimes creating separate models for sub-groups is challenging due to the requisite sensitive demographic information [97]. In addition, recent studies identified a lack of biological evidence to create treatment policies by certain characteristics, such as race, for certain diseases due to the potential for discrimination [40]. Regardless of if demographic data are used to group patients in the development of separate treatment plans, sub-groups are usually used to compare effectiveness of recommended treatment policies between patients [94]. The tradeoff between statistical advantage of using large patient subgroups for treatment plans against policies for smaller, more individualized subgroups has been considered in clinical trial settings [17]. However, the definition of subgroups in medical studies can be expanded when more data are available.

3.2 Data Management in Statistical Learning to Reduce Bias

As discussed in Section 2.3, retrospective data can be leveraged to build advanced statistical models that can guide treatment planning. However, as expanded in Sections 2.4 and 3.1, bias in current medical practices can affect the recorded EHR data [96]. Therefore, it is imperative to account for data issues before developing models to ensure that bias does not propagate into the resulting treatment plans [1, 75]. In the literature, various data management techniques have been introduced that can be exploited to address bias in advanced statistical models. In the remainder, we

first highlight how bias in data can affect treatment plans and consequently, we expand on data management techniques that can be used to address them.

One potential source for bias in EHRs is from imputation of missing data [38]. In medical data, some observations are recorded almost constantly such as from a heart rate monitor. Other features, such as medical testing, are not available at as frequent time intervals. Consequently, EHRs do not have data points for every observation at every time [7]. To use these features in statistical learning, the missing observations are often filled in, or imputed in the dataset based on different estimates of missing values [13]. Practical considerations for how data are missing influence the choice of imputation methods and affects the potential introduction of bias [89]. Note that as an alternative to imputing data, records in EHR data with missing observations may be removed. However, this may negatively impact sample size, and in some cases, has been shown to reduce model performance compared with using imputed data [38]. Also, removal of incomplete records could favor patients with comprehensive observations (partly because of their better access to more resources) [9]. EHRs do often contain a high rate of missing data; as such, the specific methods used to impute individual feature observations should be done in a way to best represent the missing data to prevent bias.

The more random the distribution and occurrences of missing (feature) observations, the less likely it is that data imputation can result in a systematic bias [89]. Generally, missing data can be categorized as **missing at random (MAR)**, **missing completely at random (MCAR)**, or **missing not at random (MNAR)**, each of which should be considered differently for imputation [37]. MAR and MCAR values can usually be estimated without introducing bias by comparing different imputation methods and choosing the method that best represents the missing data [68].

Most imputation methods estimate values based on the assumption that the values are missing randomly; however, sometimes features are MNAR and are instead caused by external non-computable factor(s) [38]. Therefore, if care is not taken, imputing MNAR values can potentially introduce bias. Unfortunately, these features are hard to identify and require a comprehensive understanding of the clinical setting where data were recorded. Indeed, there is no current solution to the bias caused by MNAR values, and hence, increased instances of missing values in EHRs has the potential to exacerbate the risk of bias in statistical models developed with imputed data. As such, it is imperative to acknowledge potential limitations of statistical studies when dealing with missing data, especially of type MNAR, and discuss the best and worse case models when these values are imputed to account for potential bias [38].

Another key data-related factor that can lead to bias in RL models learned from EHR data is sample representation. Classification models have shown to present bias when trained on imbalanced datasets (i.e., where samples are unequally represented in the dataset) [49]. Specifically, if not adjusted for, models can favor the more represented group [111]. As an example, ML models developed to predict mortality in the ICU have been shown to present differences in accuracy among ethnic groups when trained on imbalanced data [15]. Such bias can often be identified by comparing error rates across groups similar to sub-group analysis.

To alleviate introducing bias to the model due to sample representation issues, data sampling techniques can be successfully used [41]. That is, instead of training the model on imbalanced datasets, balanced sub-datasets are curated through oversampling and/or undersampling techniques. For instance, recent studies consider the number of samples from each group and the impact of a skewed population for face classification [49]. This work shows that when resampling data to contain equal samples from each race, the trained model was more accurate in classifying race, gender, and age from images. Resampling has been used outside of facial recognition, notably in healthcare, and its use has improved detection accuracy of medical events using EHR data [110]. For instance, oversampling instances of rare medication administration errors in medical datasets

for the purpose of detection and classification of such events has shown model performance improvements.

Resampling healthcare data can be done in two general ways. Minority oversampling is performed by defining some subgroup, then using the existing distributions of variables in that subgroup to either choose more of these samples in training or to generate synthetic data for this group [14]. Oversampling can increase the representation of less commonly seen patients or health states, and force treatment models to put higher "value" on their corresponding rewards during policy training. Majority undersampling is the opposite of this process; where some observations from a majority group are ignored to balance the data between groups [86]. A combination of these oversampling and undersampling techniques can be used as needed to create balanced datasets for RL.

Performing over- and undersampling is not difficult; what is challenging is to identify subgroups to resample, especially with high-dimensional feature subsets [35]. This is complicated because the stratifying features are not always even included in EHRs [76]. Hence, properly defining the sub-groups to balance the data can be quite challenging.

Defining the sub-groups can be further complicated because the data required for treatment planning is generally temporal. Mechanistically, there exist sampling methods to combat high-dimensional imbalanced time-series data that could be used for EHR data [111]. This method generates samples that preserve the correlation between features within minority classes which are established via high-dimensional density ratio-based shared nearest neighbors clustering. While this method has been established using time series data, it was proposed as a classification problem and its adoption in RL practices and healthcare still needs to be explored.

Future data management techniques may go beyond simple classification of sub-groups to represent patient populations by expanding sub-group definitions to more than just ethnic, age, or gender [1]. Systematic racism and sexism are major societal issues and are key components to understanding discrimination in medicine. However, these descriptors are not the complete definitions of sub-groups or minorities in this environment. Rather, sub-groups can be defined by multiple variables to represent a patient's state in high-dimensional latent space from EHR data [102]. Meaning that clinical feature observations can also differentiate populations beyond basic characteristics [66]. However, there is always a balance between defining more granular sub-groups and data availability. This is because training a model on a too small of a sub-group increases the potential risk of model overfitting, not allowing the learned model to properly generalize to future patients that be assigned to the sub-group [64]. As the volume of data available in EHRs increases, statistical uncertainty and the need for resampling or imputation decreases. Hence, EHRs can potentially allow for the establishment of more comprehensive sub-group definitions with sufficient observations to reduce bias in treatment planning [1].

3.3 Algorithm Design to Reduce Bias

While data management methods aim to mitigate bias prior to learning a statistical model, there exists alternative and/or complementary techniques [35] to reduce bias in treatment planning [74]. Statistical learning model designs can be instrumental where other techniques may fail to address bias towards certain patients [35]. Unlike with ML classification tasks where differences in sub-group accuracies can be easily shown, it is oftentimes difficult to represent bias in the decision making of temporal RL policies [41]. In the following, we describe ML learning methods that can be applicable in RL as well as specific RL techniques that can be potentially leveraged to mitigate bias when training a treatment planning model.

In ML, algorithms generally attempt to minimize the average error, or improve prediction accuracy, which converges to the minimum error for the majority population [95]. *Cost-sensitive* methods apply a higher penalty to errors incurred in minority groups and are common in

ML studies for imbalanced classification [11, 46]. Most related to temporal RL problems, these methods have been implemented in time-series classification [25].

Cost-sensitivity has also been used when training RL policies. For example, Ge et al. [22] use fairness-constrained policy optimization to ensure equal performance for dynamic recommendations. Here, the increased cost of mistakes on minority samples results in RL policies that exhibit parity across groups. Recently, cost-sensitive learning has been successfully applied to reward design in RL for different group samples in an imbalanced dataset [58]. Specifically, this imbalanced classification MDP framework changes the reward function so that samples from minority groups have larger positive and negative values for rewards and penalties, respectively, compared with those that are from majority groups. Therefore, the reward function is optimized in a way that the resulting policy is not overly influenced by samples from majority groups despite their higher representation.

Note that in general, applying cost-sensitive RL requires some method of identification of subgroups, which can be difficult with high-dimensional patient data. For this reason, cost-sensitivity is usually applied to ensure fairness between easily identifiable demographic groups. There are, however, successful instances that use cost-sensitive RL on multiple sub-groups defined using different or multiple features to ensure that rewards are proportionate to that group's presence in the training data [105]. High-dimensional clustering of patient data has shown to improve personalized policy performance when sufficient samples exist for each cluster [18]. The advancement beyond binary classification in patients greatly increases the potential effectiveness of bias management in cost-sensitive learning.

Additionally, models can be trained to ensure equal treatment performance, asset allocation, and outcomes between patient groups to increase fairness in treatment policies. For instance, Rajkomar et al. [80] recommends to define protected sub-groups by basic patient characteristics subject to potential discrimination. Once patients are grouped, models are trained with respect to different fairness goals while ensuring that outcomes are consistent among the sub-groups. Similar fairness constraints in RL models have shown promise in non-healthcare applications. For instance, Wen et al. [99] considers approving loan applications and shows that fairness constraints reduce biased decision-making. Another study [23] develops a multi-objective RL framework to balance the tradeoff between fairness and accuracy in RL policies for recommender systems. Such frameworks could be adapted to the healthcare domain to incorporate equitable rewards for subgroups as an objective. Introducing model constraints for fairness requires researchers to group patients based on potentially sensitive demographic or medical information. Hence, the aforementioned methods for identifying sub-groups who may be underrepresented in the data [18, 105] can pave the way to creating clinically relevant groups to be protected by fairness requirements.

Given adequate data and the information within them, RL treatment policies can be learned rather well for patients in the majority; however, these exact policies do not necessarily generalize well to underrepresented patients as their characteristics may be different [5]. Note that learning policies directly from such underrepresented patients run the risk of overfitting due to the small sample size and may not necessarily provide a good alternative [108]. Hence, different techniques can be used to adapt the learned policies from the majority to such underrepresented patients. For instance, Baucum et al. [5] learns a policy for the majority of patients, and carefully adjusts this policy using variational inference for underrepresented patients who may respond differently to treatments. The results show that (noisy) Bayesian policy updates can identify effective policies for underrepresented patients.

Transfer learning in RL can also be used to adapt policies to underrepresented patients. Transfer learning is the general practice of incorporating outside information to accelerate learning, particularly when live testing and data collection are not possible [112]. Transfer learning has been

successfully used for bias management in healthcare applications. For instance, Parbhoo et al. [73] learns policies from data-rich European HIV patient information. Then, using clustered latent representations of patients, policy knowledge is transferred to effectively treat African HIV patients, from whom less data are available.

Some learning techniques alter the RL framework itself to adapt policies to new environments. **Hidden Parameter Markov models (HiP-MDPs)** can be used to transfer policy information to similar domains [104]. HiP-MDPs learn underlying global parameters using a latent representation of the data to efficiently transfer policy information to new data during testing [106]. Meta-RL is another technique designed to leverage prior experiences to quickly adapt to similar tasks. It is currently used to improve autonomous vehicle driving for different drivers and traffic environments [109]. Hence, such RL frameworks that effectively use limited data could be further explored in treatment planning for underrepresented groups in large EHRs.

Reward shaping is another technique that incorporates domain knowledge into rewards to assist learning in RL [36]. Specifically, reward shaping allows for guiding how an optimal policy is learned by introducing short-term rewards to the reward function rather than solely relying on long-term rewards [29]. In this way, insights from clinicians can influence RL by providing feedback when observed rewards in EHRs may be sparse, such as when dealing with infrequent test results. Although the existing literature has not specifically exploited reward shaping to manage and mitigate bias, such extension seems an immediate next step. In this way, primary rewards are still related to long-term patient health, but simple, short-term rewards can incorporate fairness goals based on domain knowledge from clinicians [32, 69].

Conservative methods for reward estimation can also be used to reduce bias in RL policies. Limiting the value of estimated rewards in off-line settings has shown promise in risk-averse RL, particularly in robotics [47, 48]. Estimating rewards with a pessimistic approach, particularly for new state-action pairs, prevents overestimation of policy effectiveness when applying RL to the real world [24, 54]. Conservative reward estimates have shown to create safer policies for robotics simulation environments [16, 31]. Such an approach could be extended to healthcare applications to prevent RL models from overestimating the effect of treatment policies in new patients. Other conservative off-line learning approaches limit policy actions to those similar in the data to prevent actions where there are high levels of uncertainty [91]. RL methods that prevent off-line estimation bias can be leveraged in the future to prevent potential bias due to the domain shift in EHRs and real patients.

3.4 Clinical Implementation of Treatment Planning Models to Reduce Bias

Explainable and understandable treatment plans can allow clinicians to manage potential bias especially if it is not previously mitigated using data management and algorithmic design techniques [107]. Such treatment plans can reduce perceived and actual uncertainty of bias [34]. Here, we discuss how various concepts can be leveraged to further explain RL treatment plans and improve the understanding of potential bias.

As discussed in Section 2.4, explaining decision-making in RL policies is a challenge. Interdisciplinary collaboration between those who develop RL models and domain experts has been suggested to improve explainability [33, 77]. Therefore, when creating RL treatment policies, clinicians should be actively involved to provide insights and to understand how policies are learned.

Clinical considerations can be incorporated into learning RL policies to improve model development [35]. For instance, imposing certain constraints on the model can ensure that recommended actions meet pre-defined criteria set by clinicians [63]. This allows for recommendations to be made based on high-dimensional data, yet clinical criteria still be met in an understandable way.

Table 1. Summary of Related Works Reviewed and their Particular Areas of Contribution and Relevance to Bias Mitigation

Bias Management Method	References
Missing data	[13, 21, 30, 37, 38, 68, 89]
Resampling	[1, 15, 30, 49, 71, 74, 80, 86, 95, 110, 111]
Sub-group analysis	[2, 5, 9, 10, 17, 18, 22, 32, 35, 45, 51, 66, 71, 80, 83, 88, 94, 99, 101, 102]
Cost-sensitive learning	[11, 25, 46, 110]
Transfer learning	[73, 79, 104, 106, 109]
Reward shaping	[22, 23, 29, 36, 54, 58, 63, 69, 99, 105]
Domain shift	[5, 16, 48, 54, 91],
Explainability	[1, 21, 33, 41, 53, 61, 63, 69, 74, 77, 83, 93, 96]

Hence, clinical criteria that promote equality can be incorporated into the model as constraints to achieve bias mitigation.

A clinician-in-the-loop framework, which uses patient health states to recommend treatment options to clinicians, rather than implementing the actions directly, can be further used to discourage bias [93]. This framework is useful for multiple reasons. Primarily, there are ethical concerns with automatic administration of any medicine [19]. In addition, patient health reward structures are hard to fully define [93]. Because patient health rewards are complex, algorithms may suggest multiple near-optimal options for the clinician to choose from when administering treatment. This framework allows clinicians to use information such as patient preference or pain tolerance that is not available in EHRs when making decisions. In addition, this framework facilitates online learning as actions suggested by RL policies are used to collect further data while alleviating some ethical concerns of applying learned treatment plans. Table 1 summarizes the related works reviewed and their particular areas of contribution and relevance to bias mitigation.

4 DISCUSSION

This study reviews the existing literature relating to bias and bias management when applying RL for treatment planning. RL can effectively learn from retrospective EHR data to create treatment policies [55, 56, 87]. Although temporal EHR data lend themselves well to RL, the potential for bias is a major challenge when implementing RL systems. Notably, clinical and non-clinical influences affect the data collected in EHRs and result in sample representation issues and increased missing data [7, 41].

Through a comprehensive survey of the literature, this review highlights past and ongoing efforts to mitigate bias in RL when applied to treatment planning. The major themes that emerge relate to (1) various strategies to better manage the data from which treatment policies are learned; (2) altering the design of learning algorithms to ensure that they are learned in an equitable manner; and (3) making sure that the resulting policies are explainable and easily understood in a clinical setting.

Data can be managed to reduce bias in treatment recommendations [14, 111]. Two key challenges in data that can lead to bias are data missingness and lack of sample representation. Missing values are often imputed using different techniques and based on their missingness type. However, care must be taken during imputation not to introduce potential bias. For problems with a lack of sample representation in datasets, minority oversampling and/or majority undersampling data can be used [44, 86]. However, these techniques are not panaceas as they still require identification of appropriate sub-groups from high-dimensional EHR data [80].

Further, algorithmic design can be leveraged to mitigate bias when learning treatment policies [57]. Cost-sensitive methods can be used to ensure that learning remains sensitive with respective minority sub-groups [25]. Despite the exploration of this concept within RL, much of past work is grounded in classification where the ground truth is available and minority groups are easily identifiable. Hence, there exists a gap in research on using cost-sensitivity in reward design for developing temporal policies. Similarly, reward shaping could potentially be used to incorporate fairness goals based on domain knowledge from clinicians. Additionally, different techniques currently exist for adapting policies for new circumstances. This includes variational and Bayesian inference, transfer learning, HiP-MDPs, and meta-RL [5, 104, 106, 112]. However, more work is needed to further these techniques and ensure that they produce quality policies for any underrepresented patient with their own unique characteristics. Conservative reward estimation methods also have shown to reduce bias in off-line RL [24, 54]. However, further research is needed to investigate the extent to which implementing and potentially combining this approach with other techniques mitigates bias when learning treatment plans.

Lastly, RL policies need to be explainable and understandable to gain patient and physician buy-in and to ensure that no obvious bias is introduced during learning. Unfortunately, due to the temporal nature of RL policies and complexity of patient health data, explaining the policies and identifying potential bias can be difficult [77]. However, imposing certain criteria as constraints and involving physicians as part of model development efforts have the potential to lead to more explainable, understandable, and ultimately equitable treatment plans. A clinician-in-the-loop framework that allows for investigating near-optimal policies and selecting one that most closely matches clinical intuition is another way to improve explainability of the policies and their potential fairness.

5 CONCLUSION

This study reviews the literature on bias and bias management in RL when applied to treatment planning and identifies major themes that can be used to mitigate bias. Despite the growing literature on bias and bias management in RL when applied to health care, there still is considerable work to be done to improve bias management in RL. Future work should focus on specific methods to identify bias in RL policies and to mitigate it, especially when RL is applied to retrospectively collect EHR data. Improving bias management is the key to successful implementation of RL treatment planning policies in practice.

REFERENCES

[1] Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K. Dwivedi, John D'Ambra, and K. N. Shen. 2021. Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management* 60 (2021). https://www.sciencedirect.com/science/article/pii/S0268401221000803

[2] Mohamed Alosh, Kathleen Fritsch, Mohammad Huque, Kooros Mahjoob, Gene Pennello, Mark Rothmann, Estelle Russek-Cohen, Fraser Smith, Stephen Wilson, and Lilly Yue. 2015. Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research* 7, 4 (2015), 286–303.

[3] Onur Asan and Enid Montague. 2012. Physician interactions with electronic health records in primary care. *Health Systems* 1, 2 (2012), 96–103.

[4] Matthew Baucum, Anahita Khojandi, and Rama Vasudevan. 2021a. Improving deep reinforcement learning with transitional variational autoencoders: A healthcare application. *IEEE Journal of Biomedical and Health Informatics* 25, 6 (2021), 2273–2280. https://doi.org/10.1109/JBHI.2020.3027443

[5] Matt Baucum, Anahita Khojandi, Rama Vasudevan, and Robert Davis. 2022. Adapting reinforcement learning treatment policies using limited data to personalize critical care. *INFORMS Journal on Data Science* (2022).

[6] Matt Baucum, Anahita Khojandi, Rama Vasudevan, and Ritesh Ramdhani. 2023. Optimizing patient-specific medication regimen policies using wearable sensors in parkinson's disease. *Management Science* 0, 0 (2023).

[7] Brett K. Beaulieu-Jones, Jason H. Moore, and Pooled Resource Open-Access ALS Clinical Trials Consortium. 2017. Missing data imputation in the electronic health record using deeply learned autoencoders. In *Proceedings of the Pacific Symposium on Biocomputing 2017*. World Scientific, 207–218.

[8] Hillary Bekker. 2015. Using decision making theory to inform clinical practice. *Shared Decision Making in Healthcare: Achieving Evidence-based Patient Choice* (2015).

[9] Susannah M. Bernheim, Joseph S. Ross, Harlan M. Krumholz, and Elizabeth H. Bradley. 2008. Influence of patients' socioeconomic status on clinical management decisions: A qualitative study. *The Annals of Family Medicine* 6, 1 (2008), 53–59.

[10] Matthew Bond, Ann Bowling, Dorothy McKee, Marian Kennelly, Adrian P. Banning, Nigel Dudley, Andrew Elder, and Anthony Martin. 2003. Does ageism affect the management of ischaemic heart disease? *Journal of Health Services Research & Policy* 8, 1 (2003), 40–47.

[11] Jason Brownlee. 2020. Cost-sensitive learning for imbalanced classification. *Machine Learning Mastery* (Jan 2020). https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/

[12] Pascale Carayon, Tosha B. Wetterneck, A. Joy Rivera-Rodriguez, Ann Schoofs Hundt, Peter Hoonakker, Richard Holden, and Ayse P. Gurses. 2014. Human factors systems approach to healthcare quality and patient safety. *Applied Ergonomics* 45, 1 (2014), 14–25.

[13] Changgee Chang, Yi Deng, Xiaoqian Jiang, and Qi Long. 2020. Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature Communications* 11, 1 (2020), 1–11.

[14] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[15] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? (Dec 2018). https://arxiv.org/abs/1805.12002

[16] Dogan C. Cicek, Enes Duran, Baturay Saglam, Kagan Kaya, Furkan Mutlu, and Suleyman S. Kozat. 2021. AWD3: Dynamic reduction of the estimation bias. In *Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI'21)*. IEEE, 775–779.

[17] David I. Cook, Val J. Gebski, and Anthony C. Keech. 2004. Subgroup analysis in clinical trials. *Medical Journal of Australia* 180, 6 (2004), 289.

[18] Ali el Hassouni, Mark Hoogendoorn, Martijn van Otterlo, and Eduardo Barbaro. 2018. Personalization of health interventions using cluster-based reinforcement learning. In *Proceedings of PRIMA*.

[19] Tomás Escobar-Rodríguez, Pedro Monge-Lozano, and Ma Mercedes Romero-Alonso. 2012. Acceptance of e-prescriptions and automated medication-management systems in hospitals: An extension of the technology acceptance model. *Journal of Information Systems* 26, 1 (2012), 77–96.

[20] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. 2018. An introduction to deep reinforcement learning. *CoRR* abs/1811.12560 (2018). arXiv:1811.12560 http://arxiv.org/abs/1811.12560

[21] Michael F. Furukawa, T. S. Raghu, and Benjamin B. M. Shao. 2010. Electronic medical records, nurse staffing, and nurse-sensitive patient outcomes: Evidence from California hospitals, 1998–2007. *Health Services Research* 45, 4 (2010), 941–962.

[22] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

[23] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 316–324.

[24] Xinyang Geng, Kevin Li, Abhishek Gupta, Aviral Kumar, and Sergey Levine. [n.d.]. Effective offline RL needs going beyond pessimism: Representations and distributional shift. In *Proceedings of the Decision Awareness in Reinforcement Learning Workshop at ICML 2022*.

[25] Yue Geng and Xinyu Luo. 2018. Cost-Sensitive Convolution based Neural Networks for Imbalanced Time-Series Classification. (2018). arXiv:cs.LG/1801.04396

[26] Mohammad Ghassemi, Stefan Richter, Ifeoma Eche, Tszyi Chen, John Danziger, and Leo Celi. 2014. A data-driven approach to optimized medication dosing: A focus on heparin. *Intensive Care Medicine* 40 (08 2014). https://doi.org/10.1007/s00134-014-3406-5

[27] Rachel Gold, Erika Cottrell, Arwen Bunce, Mary Middendorf, Celine Hollombe, Stuart Cowburn, Peter Mahr, and Gerardo Melgar. 2017. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *The Journal of the American Board of Family Medicine* 30, 4 (2017), 428–447.

[28] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[29] Marek Grześ. 2017. Reward shaping in episodic reinforcement learning. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS'17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 565–573.

[30] F. M. Hajjaj, M. S. Salek, M. K. A. Basra, and A. Y. Finlay. 2010. Non-clinical influences on clinical decision-making: A major challenge to evidence-based practice. *Journal of the Royal Society of Medicine* 103, 5 (May 2010), 178–87. https://doi.org/10.1258/jrsm.2010.100104

[31] Qiang He and Xinwen Hou. 2020. WD3: Taming the estimation bias in deep reinforcement learning. In *Proceedings of the 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI'20)*. IEEE, 391–398.

[32] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. 2017. Calibration for the (Computationally-Identifiable) Masses. *CoRR* abs/1711.08513 (2017). arXiv:1711.08513 http://arxiv.org/abs/1711.08513

[33] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz Rodríguez. 2020. Explainability in deep reinforcement learning. *CoRR* abs/2008.06693 (2020). arXiv:2008.06693 https://arxiv.org/abs/2008.06693

[34] Calvin WL Ho, Joseph Ali, and Karel Caals. 2020. Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. *Bulletin of the World Health Organization* 98, 4 (2020), 263.

[35] Sara Hooker. 2021. Moving Beyond "Algorithmic Bias is a Data Problem". (Apr 2021). https://www.sciencedirect.com/science/article/pii/S2666389921000611

[36] Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. 2020. Learning to utilize shaping rewards: A new approach of reward shaping. *CoRR* abs/2011.02669 (2020). arXiv:2011.02669 https://arxiv.org/abs/2011.02669

[37] Zhen Hu, Genevieve B. Melton, Elliot G. Arsoniadis, Yan Wang, Mary R. Kwaan, and Gyorgy J. Simon. 2017. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics* 68 (2017), 112–120.

[38] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC Medical Research Methodology* 17, 1 (2017), 1–10.

[39] Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III Clinical Database. (Sept 2016). https://physionet.org/content/mimiciii/1.4/

[40] Lynn B. Jorde and Stephen P. Wooding. 2004. Genetic variation, classification and 'race'. *Nature Genetics* 36, 11 (2004), S28–S33.

[41] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key Challenges for Delivering Clinical Impact with Artificial Intelligence. (Oct 2019). https://link.springer.com/article/10.1186/s12916-019-1426-2

[42] Anahita Khojandi, Lisa M. Maillart, Oleg A. Prokopyev, Mark S. Roberts, Timothy Brown, and William W. Barrington. 2014. Optimal implantable cardioverter defibrillator (ICD) generator replacement. *INFORMS Journal on Computing* 26, 3 (2014), 599–615.

[43] Anahita Khojandi, Lisa M. Maillart, Oleg A. Prokopyev, Mark S. Roberts, and Samir F. Saba. 2018. Dynamic abandon/extract decisions for failed cardiac leads. *Management Science* 64, 2 (2018), 633–651.

[44] Aki Koivu, Mikko Sairanen, Antti Airola, and Tapio Pahikkala. 2020. Synthetic minority oversampling of vital statistics data with generative adversarial networks. *Journal of the American Medical Informatics Association* 27, 11 (2020), 1667–1674.

[45] Noemi Kreif, Richard Grieve, Rosalba Radice, Zia Sadique, Roland Ramsahai, and Jasjeet S. Sekhon. 2012. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making* 32, 6 (2012), 750–763.

[46] Matjaz Kukar and Igor Kononenko. 1998. Cost-sensitive learning with neural networks.. In *ECAI*, Vol. 15. Citeseer, 88–94.

[47] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems* 32 (2019).

[48] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 1179–1191.

[49] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. (2019). arXiv:cs.CV/1908.04913

[50] Isotta Landi, Benjamin S. Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T. Dudley, Cesare Furlanello, and Riccardo Miotto. 2020. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digital Medicine* 3, 1 (2020), 1–11.

[51] Catherine R. Lesko, Nicholas C. Henderson, and Ravi Varadhan. 2018. Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology* 100 (2018), 22–31.

[52] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR* abs/2005.01643 (2020). arXiv:2005.01643 https://arxiv.org/abs/2005.01643
[53] Ping Li, Zi yan Cheng, and Gui lin Liu. 2020. Availability bias causes misdiagnoses by physicians: Direct evidence from a randomized controlled trial. *Internal Medicine* 59, 24 (2020), 3141–3146.
[54] Qing Li, Wengang Zhou, Zhenbo Lu, and Houqiang Li. 2022. Simultaneous double Q-learning with conservative advantage learning for actor-critic methods. *arXiv preprint arXiv:2205.03819* (2022).
[55] Tian-Hao Li, Zhi-Shun Wang, Wei Lu, Qian Zhang, and Deng-Feng Li. 2021. Electronic health records based reinforcement learning for treatment optimizing. *Information Systems* (2021), 101878.
[56] Yuxi Li. 2017. Deep reinforcement learning: An overview. *CoRR* abs/1701.07274 (2017). http://dblp.uni-trier.de/db/journals/corr/corr1701.html#Li17b
[57] Enlu Lin, Qiong Chen, and Xiaoming Qi. 2019. Deep Reinforcement Learning for Imbalanced Classification. (2019). arXiv:cs.LG/1901.01379
[58] Enlu Lin, Qiong Chen, and Xiaoming Qi. 2019. Deep reinforcement learning for imbalanced classification. *CoRR* abs/1901.01379 (2019). arXiv:1901.01379 http://arxiv.org/abs/1901.01379
[59] Michael L. Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*. Morgan-Kaufmann, 157–163.
[60] Ning Liu, Ying Liu, Brent Logan, Zhiyuan Xu, Jian Tang, and Yanzhi Wang. 2018. Deep reinforcement learning for dynamic treatment regimes on medical registry data. *CoRR* abs/1801.09271 (2018). arXiv:1801.09271 http://arxiv.org/abs/1801.09271
[61] S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng. 2020. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of Medical Internet Research* 22, 7 (2020), e18477. https://doi.org/10.2196/18477
[62] Zeyu Liu, Anahita Khojandi, Xueping Li, Akram Mohammed, Robert L. Davis, and Rishikesan Kamaleswaran. 2022. A machine learning–enabled partially observable markov decision process framework for early sepsis prediction. *INFORMS J. on Computing* 34, 4 (July-August 2022), 2039–2057. https://doi.org/10.1287/ijoc.2022.1176
[63] MingYu Lu, Zachary Shahn, Daby Sow, Finale Doshi-Velez, and Li-wei H Lehman. 2020. Is deep reinforcement learning ready for practical applications in healthcare? A sensitivity analysis of duel-DDQN for hemodynamic management in sepsis patients. In *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association, 773.
[64] Aaron J. Masino, Mary Catherine Harris, Daniel Forsyth, Svetlana Ostapenko, Lakshmi Srinivasan, Christopher P. Bonafide, Fran Balamuth, Melissa Schmatz, and Robert W. Grundmeier. 2019. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PloS One* 14, 2 (2019), e0212665.
[65] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with deep reinforcement learning. *CoRR* abs/1312.5602 (2013). arXiv:1312.5602 http://arxiv.org/abs/1312.5602
[66] Susan A. Murphy. 2005. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 24, 10 (2005), 1455–1481.
[67] Shamim Nemati, Mohammad M. Ghassemi, and Gari D. Clifford. 2016. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'16)*. IEEE, 2978–2981.
[68] Cattram D. Nguyen, John B. Carlin, and Katherine J. Lee. 2021. Practical strategies for handling breakdown of multiple imputation procedures. *Emerging Themes in Epidemiology* 18, 1 (2021), 1–8.
[69] Takato Okudo and Seiji Yamada. 2021. Subgoal-based reward shaping to improve efficiency in reinforcement learning. *CoRR* abs/2104.06411 (2021). arXiv:2104.06411 https://arxiv.org/abs/2104.06411
[70] Martijn Otterlo and Marco Wiering. 2012. Reinforcement learning and Markov decision processes. *Reinforcement Learning: State of the Art* (01 2012), 3–42. https://doi.org/10.1007/978-3-642-27645-3_1
[71] Trishan Panch, Heather Mattie, and Rifat Atun. 2019. Artificial Intelligence and Algorithmic Bias: Implications for Health Systems. (Dec 2019). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875681/
[72] Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2017. Combining Kernel and Model Based Learning for HIV Therapy Selection. (Jul 2017). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543338/
[73] Sonali Parbhoo, Mario Wieser, Volker Roth, and Finale Doshi-Velez. 2020. Transfer learning from well-curated to less-resourced populations with HIV. In *Proceedings of the Machine Learning for Healthcare Conference*. PMLR, 589–609.
[74] Ravi B. Parikh, Stephanie Teeple, and Amol S. Navathe. 2019. Addressing bias in artificial intelligence in health care. *JAMA* 322, 24 (2019), 2377–2378.

[75] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Comput. Surv.* 55, 3, Article 51 (Feb 2022), 44 pages. https://doi.org/10.1145/3494672

[76] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data* 5, 1 (2018), 1–13.

[77] Erika Puiutta and Eric M. S. P. Veith. 2020. Explainable reinforcement learning: A survey. *CoRR* abs/2005.06247 (2020). arXiv:2005.06247 https://arxiv.org/abs/2005.06247

[78] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Continuous state-space models for optimal sepsis treatment - A deep reinforcement learning approach. *CoRR* abs/1705.08422 (2017). arXiv:1705.08422 http://arxiv.org/abs/1705.08422

[79] Thejan Rajapakshe, Rajib Rana, Sara Khalifa, Björn W. Schuller, and Jiajun Liu. 2021. A novel policy for pre-trained deep reinforcement learning for speech emotion recognition. *CoRR* abs/2101.00738 (2021). arXiv:2101.00738 https://arxiv.org/abs/2101.00738

[80] Alvin Rajkomar, Michael Howell, and Michaela Hardt. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. (2018). https://pubmed.ncbi.nlm.nih.gov/30508424/

[81] Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A. Oniki, Les Westberg, Calvin E. Beebe, Cui Tao, Craig G. Parker, Peter J. Haug, Stanley M. Huff, and Christopher G. Chute. 2012. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *Journal of Biomedical Informatics* 45, 4 (2012), 763–771.

[82] Elsa Riachi, Muhammad Mamdani, Michael Fralick, and Frank Rudzicz. 2021. Challenges for Reinforcement Learning in Healthcare. (2021). arXiv:cs.LG/2103.05612

[83] Wayne J. Riley. 2012. Health disparities: Gaps in access, quality and affordability of medical care. *Transactions of the American Clinical and Climatological Association* 123 (2012), 167.

[84] Patricia J. Rodriguez, Zachary J. Ward, Michael W. Long, S. Bryn Austin, and Davene R. Wright. 2021. Applied methods for estimating transition probabilities from electronic health record data. *Medical Decision Making* 41, 2 (2021), 143–152.

[85] S. Rosenbloom, William Stead, Joshua Denny, Dario Giuse, Nancy Lorenzi, Steven Brown, and Kevin Johnson. 2010. Generating clinical notes for electronic health record systems. *Applied Clinical Informatics* 1 (07 2010), 232–243. https://doi.org/10.4338/ACI-2010-03-RA-0019

[86] Fernando Sánchez-Hernández, Juan Carlos Ballesteros-Herráez, Mohamed S. Kraiem, Mercedes Sánchez-Barba, and María N. Moreno-García. 2019. Predictive modeling of ICU healthcare-associated infections from imbalanced data. Using ensembles and a clustering-based undersampling approach. *Applied Sciences* 9, 24 (2019), 5287.

[87] Andrew Schaefer and Matthew Bailey. 2005. *Modeling Medical Treatment using Markov Decision Processes.* (2005). https://link.springer.com/content/pdf/10.10072F1-4020-8066-2_23.pdf

[88] Ashkan Sharabiani, Adam Bress, Elnaz Douzali, and Houshang Darabi. 2015. Revisiting warfarin dosing using machine learning techniques. *Computational and Mathematical Methods in Medicine* 2015 (2015).

[89] Jonathan A. C. Sterne, Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 338 (2009).

[90] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press.

[91] Phillip Swazinna, Steffen Udluft, and Thomas Runkler. 2021. Overcoming model bias for robust offline deep reinforcement learning. *Engineering Applications of Artificial Intelligence* 104 (2021), 104366.

[92] Nicholas L. Syn, Andrea Li-Ann Wong, Soo-Chin Lee, Hock-Luen Teoh, James Wei Luen Yip, Raymond C. S. Seet, Wee Tiong Yeo, William Kristanto, Ping-Chong Bee, L. M. Poon, Patrick Marban, Tuck Seng Wu, Michael D. Winther, Liam R. Brunham, Richie Soong, Bee-Choo Tai, and Boon-Cher Goh. 2018. Genotype-guided versus traditional clinical dosing of warfarin in patients of Asian ancestry: A randomized controlled trial. *BMC Medicine* 16, 1 (2018), 1–10.

[93] Shengpu Tang, Aditya Modi, Michael W. Sjoding, and Jenna Wiens. 2020. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. *CoRR* abs/2007.12678 (2020). arXiv:2007.12678 https://arxiv.org/abs/2007.12678

[94] Julien Tanniou, Ingeborg Van Der Tweel, Steven Teerenstra, and Kit C. B. Roes. 2016. Subgroup analyses in confirmatory clinical trials: Time to be specific about their purposes. *BMC Medical Research Methodology* 16, 1 (2016), 1–15.

[95] Alexandra Chouldechova and Aaron Roth. 2020. *A Snapshot of the Frontiers of Fairness in Machine Learning.* (May 2020). https://dl.acm.org/doi/pdf/10.1145/3376898

[96] R. Vincent. 2014. Reinforcement learning in models of adaptive medical treatment strategies. McGill University (Canada). 2014.

[97] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine* 383, 9 (2020), 874–882.

[98] Jeremy Watts, Anahita Khojandi, Rama Vasudevan, and Ritesh Ramdhani. 2020. Optimizing individualized treatment planning for Parkinson's disease using deep reinforcement learning. In *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC'20)*. 5406–5409. https://doi.org/10.1109/EMBC44109.2020.9175311

[99] Min Wen, Osbert Bastani, and Ufuk Topcu. 2021. Algorithms for fairness in sequential decision making. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (PMLR)*, 1144–1152.

[100] Wei-Hung Weng, Mingwu Gao, Ze He, Susu Yan, and Peter Szolovits. 2017. Representation and reinforcement learning for personalized glycemic control in septic patients. *CoRR* abs/1712.00654 (2017). arXiv:1712.00654 http://arxiv.org/abs/1712.00654

[101] Jeff Whittle, Joseph Conigliaro, C. B. Good, and Richard P. Lofgren. 1993. Racial differences in the use of invasive cardiovascular procedures in the Department of Veterans Affairs medical system. *New England Journal of Medicine* 329, 9 (1993), 621–627.

[102] Edwin S. Wong, Jean Yoon, Rebecca I. Piegari, Ann-Marie M Rosland, Stephan D. Fihn, and Evelyn T. Chang. 2018. Identifying latent subgroups of high-risk patients using risk score trajectories. *Journal of General Internal Medicine* 33, 12 (2018), 2120–2126.

[103] Jionglin Wu, Jason Roy, and Walter F. Stewart. 2010. Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. *Medical Care* (2010), S106–S113.

[104] Jiachen Yang, Brenden K. Petersen, Hongyuan Zha, and Daniel M. Faissol. 2019. Single episode policy transfer in reinforcement learning. *CoRR* abs/1910.07719 (2019). arXiv:1910.07719 http://arxiv.org/abs/1910.07719

[105] Jenny Yang, Andrew A. S. Soltan, and David A. Clifton. 2022. Algorithmic fairness and bias mitigation for clinical machine learning: A new utility for deep reinforcement learning. *medRxiv* (2022).

[106] Jiayu Yao, Taylor Killian, George Konidaris, and Finale Doshi-Velez. 2018. Direct policy transfer via hidden parameter markov decision processes. In *Proceedings of the LLARLA Workshop, FAIM*, Vol. 2018.

[107] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2021. Reinforcement learning in healthcare: A survey. *ACM Comput. Surv.* 55, 1, Article 5 (Nov 2021), 36 pages. https://doi.org/10.1145/3477600

[108] Daochen Zha, Kwei-Herng Lai, Qiaoyu Tan, Sirui Ding, Na Zou, and Xia Ben Hu. 2022. Towards automated imbalanced learning with deep hierarchical reinforcement learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2476–2485.

[109] Songan Zhang, Lu Wen, Huei Peng, and H. Eric Tseng. 2021. Quick learner automated vehicle adapting its roadmanship to varying traffic cultures with meta reinforcement learning. *CoRR* abs/2104.08876 (2021). arXiv:2104.08876 https://arxiv.org/abs/2104.08876

[110] Yang Zhao, Zoie Shui-Yee Wong, and Kwok Leung Tsui. 2018. A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. (May 2018). https://www.hindawi.com/journals/jhe/2018/6275435/

[111] Tuanfei Zhu, Yaping Lin, and Yonghe Liu. 2020. Oversampling for imbalanced time series data. *CoRR* abs/2004.06373 (2020). arXiv:2004.06373 https://arxiv.org/abs/2004.06373

[112] Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. 2020. Transfer learning in deep reinforcement learning: A survey. *CoRR* abs/2009.07888 (2020). arXiv:2009.07888 https://arxiv.org/abs/2009.07888

Received 17 February 2022; revised 16 November 2022; accepted 11 July 2023