Development of Public Health Policy by Digital Twin Microsimulation and Q-learning: A COVID-19 Booster Case Study

Guoxuan Ma[1], Sicong Xie[2], Lili Zhao[3*] and Jian Kang[1*]
Department of Biostatistics, University of Michigan[1]
Department of Statistics, University of Michigan[2]
Feinberg School of Medicine, Northwestern University[3]

May 15, 2026

Abstract

The COVID-19 pandemic highlighted the urgent need for effective vaccine policies, but traditional clinical trials often lack sufficient data to capture the diverse population characteristics necessary for comprehensive public health strategies. Ethical concerns around randomized trials during a pandemic further complicate policy development for public health. Reinforcement Learning (RL) offers a promising alternative for vaccine policy development. However, direct online RL exploration in real-world scenarios can result in suboptimal and potentially harmful decisions. This study proposes a novel framework combining tabular Q-learning with microsimulation, where a Recurrent Neural Network (RNN) serves as a digital twin environment simulator of the target population. This digital twin captures temporal associations between infection and patient characteristics to generate realistic individual disease trajectories, enabling safe and efficient policy learning without real-world interaction. Our tabular Q-learning model produces an interpretable policy table that balances the risks of severe infection against vaccination side effects. Applied to COVID-19 booster policies, the learned Q-learning-based policy outperforms current practices, offering a path toward more effective vaccination strategies.

Keywords: Vaccine Policy, Public Health, Q-learning, Recurrent Neural Network, COVID-19 Vaccine Booster

*To whom correspondence should be addressed: zhaolili@northwestern.edu and jiankang@umich.edu

1 Introduction

The COVID-19 pandemic underscored the critical importance of rapid and effective vaccination strategies to control the spread of the virus and minimize the burden on healthcare systems. However, developing optimal vaccination policies for public health is fraught with challenges. Clinical trials do not have sufficient data to evaluate vaccine policies comprehensively, as they often enroll subjects with specific characteristics that may not be representative of the general population (Jüni et al., 2001; Lander et al., 2019). As a result, the findings from these trials may not always generalize well to the broader population, potentially hindering the development of comprehensive vaccination policies. For instance, people taking immunosuppressant medications were excluded from the trials developing BNT162b2 (Pfizer-BioNTech) and mRNA-1273 (Moderna) vaccines (Polack et al., 2020; Baden et al., 2021). The lack of data in this group regarding the vaccines efficacy has prevented the development of an effective vaccine policy against COVID-19 infections for immunosuppressed patients (Risk et al., 2022a). Moreover, conducting large-scale randomized trials on vaccine policy evaluation during a pandemic poses significant ethical challenges. Randomizing subjects to the group that do not receive the vaccine can place participants at increased risk of COVID-19 infection, raising ethical concerns about exposing groups of people to potential harm (Adebamowo et al., 2014; Monrad, 2020). These challenges highlight the need for alternative approaches to develop and improve vaccine policies using existing real-world data while adhering to safety and ethical standards.

The electronic health record (EHR) dataset from a medical institution, which spans both the pre-pandemic and pandemic periods, provides a comprehensive resource for studying COVID-19 vaccination strategies in a real-world context. This dataset comprises extensive health records from 1,224,147 patients, including demographics (e.g., age, gender, and race), COVID-19 vaccination dates, infection history, and other baseline characteristics such as

immunosuppressant usage, number of hospital visits, and comorbidities. The breadth and granularity of this dataset enable a systematic investigation of vaccination policies that account for heterogeneity across patient populations. Based on this EHR dataset, we aim to develop vaccination policies that improve upon current observed practices and better address the needs of diverse patient populations, including the immunosuppressant groups that were underrepresented in clinical trials (Polack et al., 2020; Risk et al., 2022a).

Reinforcement learning (RL) offers a promising framework for developing adaptive vaccination strategies and has been successfully applied in various fields in healthcare (Sutton and Barto, 2018; Yu et al., 2021), including cancer treatment, glucose regulation, HIV treatment, and mental diseases intervention (Tseng et al., 2017; Sun et al., 2018; Yu et al., 2019; Laber et al., 2014), but it has not been widely adopted in public health (Weltz et al., 2022). In an RL setup, an agent selects actions based on its current state, receiving feedback (rewards) and the new state from the environment. The objective is for the agent to learn an optimal policy, which is a mapping from states to actions, that maximizes the cumulative reward over time. RL is particularly suited for systems with inherent delays, where decisions are made sequentially without immediate feedback and are evaluated based on long-term outcomes. This makes RL a compelling approach for developing effective policies in public health, including policymaking in vaccination, where the health outcomes are often evaluated based on a prolonged period with delayed feedback (Yu et al., 2021).

However, RL agents do not receive explicit instructions on which actions to take; instead, they learn the best actions through trial and error (in the online setting) or learn from the existing data (in the offline setting). While the online trial-and-error process encourages agents to explore new policies that are potentially effective, applying this approach directly in real-world scenarios can raise ethical concerns (Levine et al., 2020). Early in the training process, the trial-and-error learning mechanism often lead to suboptimal or even harmful decisions, potentially causing harm to subjects before corrective feedback

is obtained. While this may be less problematic in applying RL in Dynamic Treatment Regimes (Liu et al., 2017; Zhang, 2020; Guo et al., 2022), where treatment decisions are usually made under experts' supervision to ensure they are clinically relevant and safe, it becomes more critical in scenarios of learning vaccine policy during a global pandemic like COVID-19, where it is impossible to provide individualized supervision for the whole population. In such cases, incorrect vaccination timing or administration could lead to severe infections or serious adverse effects. This problem in online training can be resolved by an offline approach, where actions are learned based on observed data. However, the offline approach may struggle with exploring new policies that could lead to potential improvements and cannot effectively learn from the observed offline data (Levine et al., 2020). Moreover, it faces the same ethical concerns when evaluating the learned policy in real world with the absence of proper supervisions. Therefore, there is a need to develop an online RL framework based on existing data without direct interactions with the real world.

Q-learning is a Reinforcement Learning algorithm that helps agents learn how to make decisions by evaluating the potential value of different actions in various states (Watkins and Dayan, 1992). It maintains a Q-function of the state-action pairs, which represents the expected future reward of taking a particular action in a given state. To determine the best action for a given state, the agent examines the Q-values associated with all possible actions and selects the one with the highest value. The Q-value for each state-action pair is initially designed to be stored explicitly in a table, as seen in tabular Q-learning (Watkins and Dayan, 1992). Later, deep neural networks (DNN) are often used to model the Q-function for its flexibility and ability to accommodate continuous state and action spaces (Mnih et al., 2015; Liu et al., 2017; Yu et al., 2021). While deep Q-learning is useful in precision medicine or individualized treatment where state and action spaces are often large or continuous (Liu et al., 2017; Zhang, 2020; Guo et al., 2022; Wu et al., 2023), public health problems can be effectively represented with a finite and discrete set of states

and actions, policies are often applied to groups of people rather than being tailored to individual subjects. Moreover, unlike tabular Q-learning, deep Q-learning lacks theoretical convergence guarantees to the global optimum due to the complexities of neural networks and function approximation and often suffers from convergence difficulties (Watkins and Dayan, 1992; Van Hasselt et al., 2016; Fan et al., 2020).

Many efforts have focused on estimating the impact of government interventions on the epidemiological spread of COVID-19. For instance, Chernozhukov et al. (2020) utilized a causal structural model to investigate the effects of policies on COVID-19 growth and mortality rates. Eftekhari et al. (2020) compared Markovian and non-Markovian processes for lockdown allocation, while Tian et al. (2021) employed synthetic control, discontinuity regression and state-space compartmental models to evaluate intervention stringency. Tan et al. (2022) employed compartmental models to estimate transmissibility of symptomatic and asymptomatic cases. Other approaches relied on RL for government-level policy optimization for COVID-19. Kompella et al. (2020) proposed an RL-based framework for fine-grained mitigation, and Wan et al. (2021) applied RL to minimize long-term societal costs, both using epidemiological models as the environment simulator. However, these methods primarily address population-level infection dynamics or spatial spread. Because these simulators do not generate individual-level trajectories, they are unsuitable for developing vaccination policies tailored to specific subgroups. While Kerr et al. (2021) took an agent-based approach to projecting epidemic trends by simulating the state of individual people over discrete time points, their framework does not address sequential decision-making optimization required for targeted vaccination strategies.

To address these challenges, we propose a framework combining online tabular Q-learning with an RNN-based environment that interacts with the Q-learning agent. We refer to this RNN-based environment as a digital twin microsimulation model (Rutter et al., 2011; Julien et al., 2022), as it effectively simulates individual trajectories and creates a

virtual environment that mirrors real-world dynamics. Figure 1 provides an overview. To illustrate the proposed framework, we focus on the development of COVID-19 vaccine policies in Michigan, using data collected from the University of Michigan Hospital. Nonetheless, the framework itself is broadly applicable to a wide range of sequential decision-making problems in public health and can be adapted to other populations and applications. We train an RNN with Long Short-Term Memory (LSTM) architecture, which has been widely used to model sequential data and has successful applications on modeling the relationship between COVID-19 infections and vaccinations (Sherstinsky, 2020; Shen et al., 2024), to capture the complex temporal association between infection status and patients' characteristics. Then, we perform online Q-learning, where the Q-function is modeled as a table, for vaccine policy learning based on the RNN microsimulator to avoid the need of actual executing the decision on whether to receive the vaccination.

Our approach has two key contributions. First, by employing an RNN with the LSTM architecture, we create a digital twin microsimulation model as the virtual environment used in RL that can generate data for individuals that closely resembles real-world data. The RNN-based microsimulator not only avoids ethical concerns but also provides an unlimited amount of data for training, which creates "what if" scenarios, allowing us to evaluate different policies on simulated individuals with the same characteristics during the same period of time. Second, by using the tabular Q-learning, our approach produces a clear and interpretable policy table where Q-values for different actions and various groups of people can be easily read. We define the reward function by balancing the risk of severe infections and the potential side effects of the vaccination. In this paper, we focus on the vaccination policy for the COVID-19 booster dose, i.e., whether and when different groups of people should receive the booster after the second COVID-19 vaccination. The policy derived from our Q-table demonstrates superior performance compared to the current observed policy, indicating significant potential for improvement if the learned policy is adopted.

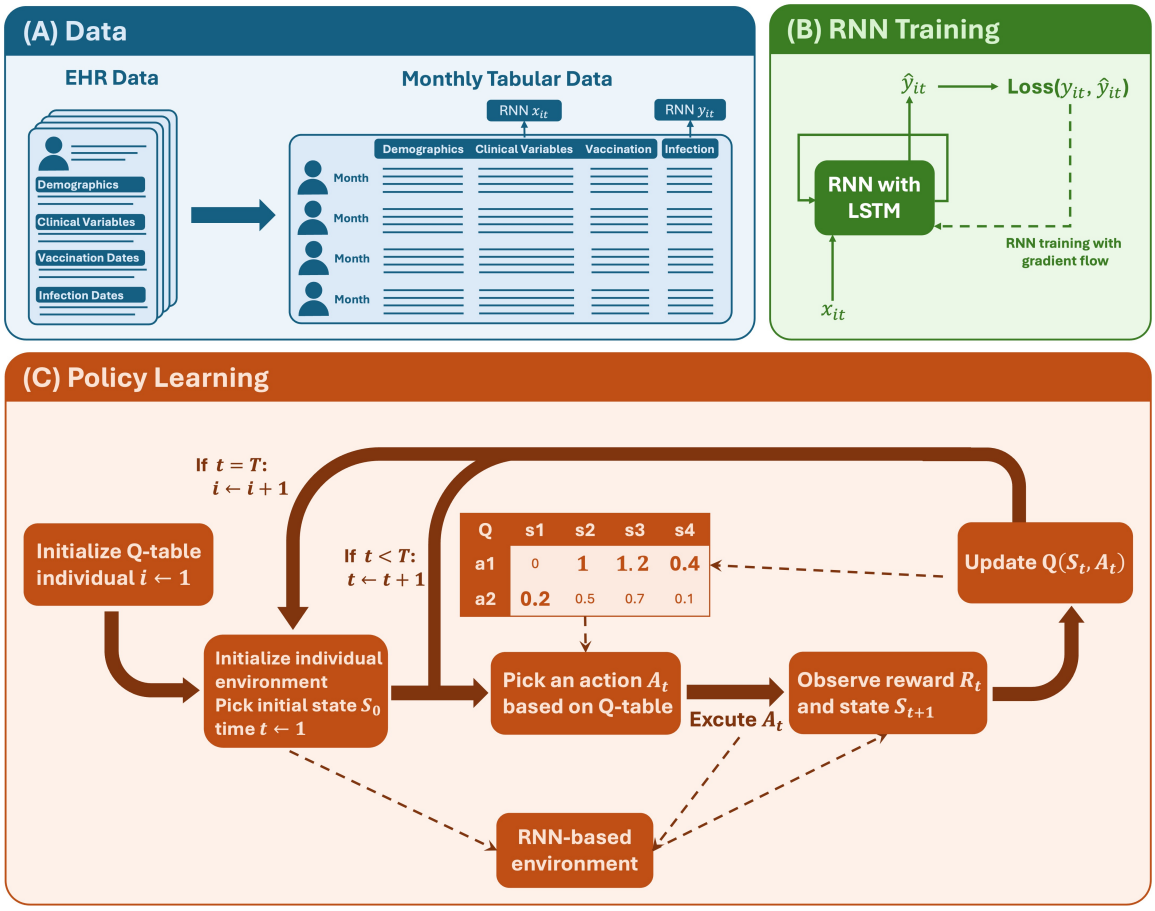


Figure 1: Our policy learning framework consisted of digital twin microsimulation and Q-learning. (A) An illustration of the data. The EHR data, including demographics, clinical variables, vaccination and infection dates, is converted to monthly tabular data for RNN training. For individual i, x_{it} represents RNN predictors and y_{it} represents RNN outcomes, for $t = 1, \ldots, T$. Section 2.4 includes details on variables included in x_{it} and y_{it}. (B) Training an RNN with LSTM architecture using monthly tabular data. (C) Steps of online tabular Q-learning where the environment is based on the fully-trained RNN with LSTM architecture from (B). Algorithm S1 provides more details on Q-learning steps.

2 Methods

2.1 Data processing

We use deidentified electronic health record (EHR) dataset from the University of Michigan Hospital and Michigan Medicine (study ID: HUM00164771). The use of the data was approved by the Institutional Review Board (IRB) at the University of Michigan. We include patients with a primary care physician and received at least one COVID-19 test at the University of Michigan Hospital. The dataset includes demographic variables, such as

age, gender, and race, as well as baseline health measures, including the number of prior hospital visits and Charlson comorbidity index (Charlson et al., 1987; Gasparini, 2018). In addition, it records time-varying variables on patients' COVID-19 infection and vaccination status. We exclude patients with race and gender missing as both variables are included in the microsimulation model. Since fewer than 1.5% patients had either variables missing, we believe the exclusion is unlikely to introduce bias. A total of 81,000 patients are included in this study. The primary outcome we consider is whether the patient has severe COVID-19 infection (requiring hospitalization), and the secondary outcome is whether the patient has general COVID-19 infection. In this study, we summarize monthly data (illustrated in Figure 1A) from the original EHR data starting from March 2020, and a patient's record terminates either at June 2022, the month of severe infection or decease, whichever comes first. The maximum span of data sequence for a patient is $T = 27$ months, from March 2020 to June 2022, with minimum of 1 month, mean of 26.6 months, and median of 27 months. Although WHO did not declare the end of COVID-19 as a global health emergency until May 2023, the pandemic had largely subsided in the United States by mid-2022 (CDC, 2024). In addition, our dataset was available only through the end of June 2022 when we started the research. For these reasons, we selected June 2022 as the termination date.

2.2 Overview

Let $\{S_t\}_{t\geq 0}$ be a Markov process with state space $\mathcal{S}$ representing an individual's baseline characteristics, vaccination history and severe infection status at each month. Let A_t be a random variable that represents the choice to receive a booster or not in month t with action space $\mathcal{A} = \{0, 1\}$. For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, denote by $P(s'|s, a) = \Pr(S_{t+1} = s' \mid S_t = s, A_t = a)$ the state transition probability function. Let $\mathcal{R}(s, a)$ be a stationary reward function of taking action a in state s. Let $R_t = \mathcal{R}(S_t, A_t)$, which is a random variable representing the reward at time t after taking action A_t at state S_t. Let s_t, a_t and r_t be the realizations of S_t, A_t and R_t respectively. Then, we model the decision-making process of

whether an individual should receive a booster at each month as a Markov Decision Process (MDP), denoted by $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, \mathcal{R}\}$. In this framework, the current state adequately summarizes the past, meaning future states are independent of past states given the current state and action. This assumption entails little loss of generality mathematically because almost any decision process can be reformulated as an MDP by appropriately aggregating the historical data (Sutton, 1997; Weltz et al., 2022).

In Reinforcement Learning (RL), an agent interacts with the environment to improve its decision-making ability over time. An agent is a decision-maker responsible for determine whether an individual should receive a COVID-19 booster shot. The agent selects an action based on the current state according to a policy $\pi : \mathcal{S} \mapsto \mathcal{A}$, which maps a patient's health status, vaccination history, and other relevant information (a state) to an action (whether to administer a booster or not). The environment, on the other hand, consists of the state transition function $P(s'|s, a)$ and the reward function $\mathcal{R}(s, a)$. The environment interacts with the agent by providing the next state and reward after each action is taken.

We first introduce the Q-learning framework to derive a good policy for booster vaccination, assuming the environment is known. However, due to ethical concerns around allowing an agent to directly interact with real-world patients in the development of booster polices, we construct a microsimulation model as a virtual environment using RNN trained on existing patient data. Algorithm S1 outlines the steps of online tabular Q-learning with the RNN-based environment in our application of booster policy development.

2.3 Booster policy learning by tabular Q-learning

We perform online tabular Q-learning given a known environment. The objective is to find a policy π^* maximizing the expected cumulative reward, defined by the value function, $V^\pi(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s_0\right]$, where $V^\pi(s_0)$ represents the expected cumulative reward starting from an initial state s_0 by following a policy π, and $\gamma \in (0, 1]$ is the discount factor on future reward. The expectation is taken over possible trajectories of the Markov process

$\{S_t\}_{t\geq 0}$ generated by the policy π starting from the initial state s_0. Because the MDP $\mathcal{M}$ is defined with a stationary transition kernel $P(s'|s,a)$, the value function is also stationary. This means for any state, its value remains constant over time, as the recursive nature of the decision-making process ensures the value depends only on the current state, not on when it is encountered. Under the MDP framework, the optimal policy π^* satisfies the Bellman equation,

$$V^*(s) = \max_{\pi} \mathbb{E}\left\{\mathcal{R}\left(s, \pi(s)\right) + \gamma V^*(s')\right\},$$

where s is the current state and s' is the next state. Instead of working directly with $V^*(s)$, Q-learning focuses on the optimal action–value function q^*, which satisfies its own Bellman optimality equation, $q^*(s,a) = \mathbb{E}\left\{\mathcal{R}(s,a) + \gamma \max_{a'} q^*(s',a')\right\}$, where $q^*(s,a)$ represents the expected cumulative reward of taking action a in state s following the optimal policy thereafter. The optimal policy is then given by $\pi^*(s) = \arg\max_a q^*(s,a)$. Since the Q-function is initially unknown, we model it as a table, where each cell (s,a) holds the estimated value of $q(s,a)$ (Watkins and Dayan, 1992). The Q-learning algorithm takes an iterative approach to updating the Q-table: after each interaction with the environment, we adjust $q(s,a)$ based on the observed reward r and the estimated future value by the following updating rule (Watkins and Dayan, 1992),

$$q(s,a) \leftarrow q(s,a) + \beta\left\{r + \gamma \max_{u} q(s',u) - q(s,a)\right\} \tag{1}$$

where β is a prespecified learning rate and s' is the next state.

In this study, we aim to learn a policy on whether and when to receive a COVID-19 booster, so we only consider subjects with at least two COVID-19 vaccinations and their trajectories after their second vaccinations. At any month t, we decide the state $S_t \in \mathcal{S}$ consists of four relevant variables: age (categorical, 18-29/30-49/50-64/65+), baseline immunosuppressant usage (binary), months since the last vaccination (categorical, 0-4/5-6/7+), and the severe infection status (binary). The action $A_t \in \mathcal{A} = \{0,1\}$ indicates

whether or not a booster is received. In this study, we follow the Centers for Disease Control and Prevention (CDC) guidelines that an additional COVID-19 vaccine should be at least 4 months following the previous dose. The guideline was for adult ages 65 years and older but we generalize it to all age groups for simplicity. Following this guideline, A_t is constrained to be 0 regardless of the values in Q-table for t within 4 months of the second vaccination.

2.4 Creating the environment through microsimulation

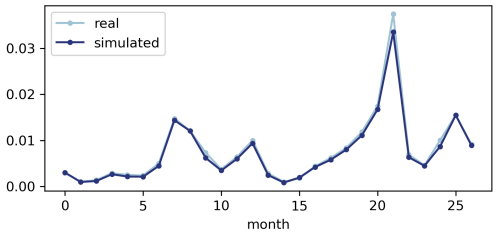
The RL environment is characterized by two key components: the state transition rules and the reward function. We simulate state transitions using RNN, enabling the microsimulation of individual trajectories. The reward function is designed to balance the risk of severe infection with potential adverse effects of a COVID-19 booster.

Microsimulation of state transitions Let W_t represent a set of additional variables relevant to the individual's profile, though of less interest to the vaccine policy decision and not included in the state variable S_t. It includes an individual's baseline and time-varying characteristics. In this study, baseline characteristics include gender (Female/Male), race (categorical, Caucasian/African American/Others), baseline number of hospital visits (categorical, 0-4/5-9/10-19/20-49/50+), and baseline Charlson comorbidity (categorical, 0/1-2/3-4/5+). The baseline period is one year before the study period starts. Time-varying variables include COVID-19 variant (categorical, Alpha/Delta/Omicron) and total number of vaccinations received up to month t (integer, 0-4).

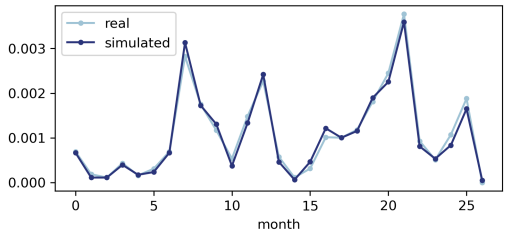
We train an RNN with LSTM units to approximate the transition dynamics between states. Conceptually, for month $t = 1, \ldots, T-1$, let $X_t = [S_t, W_t, A_t]$ denote the RNN predictors and $Y_{t+1} = [S_{t+1}, W_{t+1}]$ denote the RNN outcomes. The RNN takes as input the sequence of predictors $X_1, \ldots, X_t$ and predicts the outcome Y_{t+1}. The transition dynamics from $[S_t, W_t]$ to $[S_{t+1}, W_{t+1}]$ given action A_t are modeled by the fully trained RNN, from where we obtain the transition dynamics from S_t to S_{t+1}.

In this study, certain state variables, including age, baseline immunosuppressant usage and months since last vaccination, as well as the set of additional variables W_t are deterministic given action and time. Therefore, these variables are excluded from the outcomes, as their transitions are fixed. Additionally, since severe infections are treated as a terminal event (i.e., the trajectory is terminated upon a severe infection), it is always zero in S_t for month t before termination. As the general infection is an important variable that affects the likelihood of severe infection, we include the binary general infection status besides the binary severe infection status into the outcome variable. The output layer of the RNN uses a sigmoid activation to predict the probability of infections in the next month. This allows us to sample transitions from S_t to S_{t+1} based on the underlying state transition function, and realize microsimulation of individual trajectories in alignment with the Q-learning policy. Section 3.1 shows that the state transition probabilities estimated from the microsimulated individuals match well with those evaluated on real-data.

Define rewards We define deterministic reward in each month t, given state s_t and action a_t, $\mathcal{R}(s_t, a_t) = -I(s_t, a_t) \times (1 + \alpha \times a_t)$, where $I(s_t, a_t) : \mathcal{S} \times \mathcal{A} \to \{0, 1\}$ is the severe COVID-19 infection status for the next month, and α represents the relative cost of receiving a booster. The next-month severe COVID-19 infection status $I(s_t, a_t)$ can be sampled using the probability generated by the fully trained RNN. The reward consists of two components: the predicted risk of severe infection in the next month and the potential side events of the booster. The parameter α quantifies the perceived harm of receiving a booster relative to the risk of a severe infection. As the Q-learning seeks a policy to maximize the expected cumulative reward, when α is small, it is likely to recommend boosters for all groups due to protective effect of vaccines against severe infections. Conversely, a large α may result in a policy that discourages booster administration, as the perceived harm outweighs the infection risk. In practice, the choice of α should reflect the relative importance placed on the risk of infection versus the harm of vaccination. We discuss our choice of α in Section



(a) Marginal general infection rate by month.



(b) Marginal severe Infection rate by month.

Figure 2: Marginal general infection rate and marginal severe infection rate by month summarized from the simulated data and the real data.

3.2.1 when presenting our results.

3 Application on Learning COVID-19 Booster Policy

In this section, we first present results of our microsimulations, showing that our RNN-based environment resembles real data well in terms of both marginal infection probabilities and conditional infection probabilities. Then, we present our Q-learning results, showing that the Q-table-based policy has an advantage over other policies in terms of rewards. We interpret the Q-table-based policy on whether and when to receive the booster on different groups of population when choosing selected vaccine costs.

3.1 Microsimulation to create environment

We train the RNN using monthly data of the 81,000 patients starting from March 2020 to June 2022. We use an RNN with 2 stacked LSTM layer for training. Each LSTM layer has 128 hidden nodes with dropout rate 0.2 (Srivastava et al., 2014). During training, we use the Adam optimizer with learning rate 10^{-4} for 2,000 epochs (Kingma and Ba, 2014). Sensitivity analysis shows that mild changes in hyperparameters do not change results significantly (see Supplementary Materials S2). To evaluate the RNN-based environment, we simulate a data sequence starting from the predictors at March 2020 for each of the 81,000 patients by the trained RNN. We summarize the severe infection rate and the general infection rate from the simulated data and compare them with the real EHR data.

The simulated population marginal general infection rate and marginal severe infection

Table 1: Severe infection rate (‰) conditional on one variable: (a) age (b) number of previous COVID-19 vaccines (c) number of baseline hospital visits (d) comorbidity score.

(a) Age

	Simulated	Real
Age [0, 18)	0.67	0.67
Age [18, 30)	0.92	1.06
Age [30, 50)	1.07	1.12
Age [50, 65)	1.14	1.05
Age 65+	1.26	1.26

(b) Number of vaccines

	Simulated	Real
numVax = 0	1.27	1.28
numVax = 1	1.04	1.06
numVax = 2	0.72	0.73
numVax = 3	0.73	0.85
numVax = 4	0.32	1.00

(c) Number of visits

	Simulated	Real
numVisits [0, 5)	0.76	0.79
numVisits [5, 10)	0.74	0.75
numVisits [10, 20)	1.08	1.05
numVisits [20, 50)	1.84	1.76
numVisits 50+	2.84	3.28

(d) Comorbidity

	Simulated	Real
Comorbidity [0, 1)	0.76	0.77
Comorbidity [1, 3)	1.43	1.39
Comorbidity [3, 5)	2.18	2.24
Comorbidity 5+	3.02	3.15

rate over the 27 months are 7.25‰ and 1.06‰ respectively, which are close to the observed values in real data (7.91‰ and 1.06‰ respectively). Figure 2 shows the marginal general infection rate and marginal severe infection rate within the population at each month for the simulated data and the real data. The RNN-based environment fits very well both the marginal general infection rate and marginal severe infection rate within the population over the 27 months. The simulated marginal infection rates is very close to the real infection rates at each month.

Table 1 shows the severe infection rate conditional on one variable and Table 2 shows the severe infection rate conditional on multiple variables. The simulated conditional severe infection rates are very close to the rates observed in real data in most cases. Rarely, the simulated conditional severe infection rate has some difference with the observed values because there are limited data points within that category.

To assess whether the simulated decision process based on the trained RNN satisfies the Markov property, we use the test proposed by Shi et al. (2020) on multiple random subsets, and the low rejection rates (0%–2%) indicate that the Markov assumption is not violated

Table 2: Severe infection rate (simulated/observed, ‰) conditional on multiple variables: (a) Baseline immunosuppressant status and gender (b) COVID variant and race.

(a) Baseline immunosuppressant status and gender.

	imm_baseline = 0	**imm_baseline = 1**
gender = 0	0.95 / 0.95	1.53 / 1.51
gender = 1	1.01 / 1.05	2.02 / 1.85

(b) COVID variant and race

	Variant None	**Variant Delta**	**Variant Omicron**
Race Caucasian	0.76 / 0.76	1.56 / 1.53	0.72 / 0.79
Race African American	1.71 / 1.61	4.60 / 4.57	1.30 / 1.55
Race Others	0.75 / 0.80	1.60 / 1.69	0.71 / 0.76

(see Supplementary Materials S3 for details). These results suggest that the trained RNN provides a reasonable approximation of a Markovian environment, which justifies the use of (tabular) Q-learning to obtain an approximately optimal policy.

Overall, results show that the digital twin (RNN-based environment) is reliable for the online tabular Q-learning. The trained RNN can simulate data with very similar marginal and conditional infection rates with those in the real data. It well captures the relationship between the infection status and both the baseline and time-varying variables.

3.2 Booster policy learning

We include three variables in state $\mathcal{S}$ in the online tabular Q-learning: age, baseline immunosuppressant status, and number of months to the second COVID-19 vaccination. Since we aim to learn a policy on whether to receive a COVID-19 booster at a specific month, we only consider subjects with at least two COVID-19 vaccinations. In this study, we focus on the policy of the first booster.

3.2.1 Reward evaluation

We compare the population average rewards over months of the Q-table-based policy and three other policies: observed policy from data, all receiving a booster, and none receiving a booster. For the policy from data, we extract whether and when each subject received the first booster from the real data. For the policy of receiving a booster, we

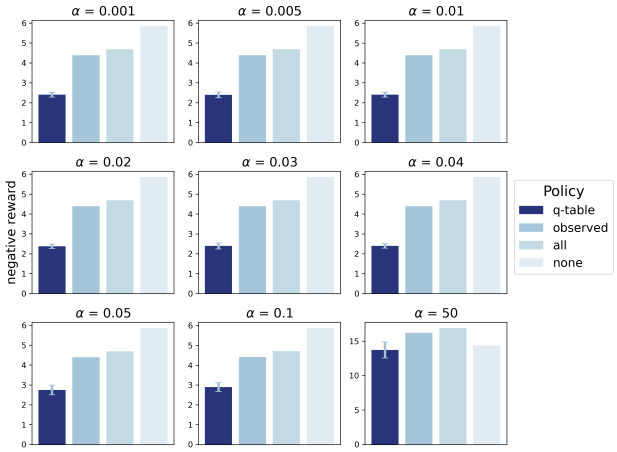


Figure 3: Negative rewards ($\times 10^{-4}$) for the Q-table-based policy (q-table), the observed policy from data (observed), the policy of all receiving a booster (all), the policy of none receiving a booster (none). The error bar represents mean $\pm$ s.d. of the population average negative rewards over the 20 replicates for the Q-table-based policy.

randomly pick a month between the 5th from the second vaccination and the last month of the study for each subject to receive a booster. For the policy of never receiving a booster, no one receives any booster at any month. We train the Q-table for 30 epochs and repeat the training 20 times with different random seeds. Supplementary Materials S1 shows that the tabular Q-learning algorithm converges before 30 epochs in our study. The discount factor γ in the tabular Q-learning is fixed at 0.99.

The vaccine cost α is an important hyperparameter that controls the belief of the relative harmness of a severe infection and the booster. One of the choice is to determine a reasonable range of vaccine cost based on mortality rates after a severe infection and the booster. We compute the mortality rate after 30 days among people who received the booster (0.04%) and the mortality rate after 30 days among people who had a severe infection (1.05%) from the real data. We use the ratio (0.04) of the two mortality rates for a proxy of the relative risk of a booster to a severe infection, so we primarily focus on the vaccine cost α around 0.04.

Figure 3 shows the negative rewards (the lower the better) for the four policies and selected vaccine costs around 0.04. For all the vaccine costs, the Q-table-based policy

consistently has the highest reward. The Q-table-based policy has a stable reward over the 20 replicates. When the vaccine cost increases (vaccine cost = 50), the Q-table-based policy gets close to the policy of never receiving a booster, as the large vaccine cost discourages it to receive a booster at any time point. It is worth-noting that the Q-table-based policy has a higher reward than the policy from the real data in all cases. This indicates the policy that was followed by the general population is sub-optimal and could have been improved by our Q-table-based policy.

Supplementary Materials S4 includes several sensitivity analyses, including varying different choices of γ, adding gender as an additional state, and refining age into more categories. Results show that our results are robust to the above variations.

In Section 4, we provide the comparison between tabular Q-learning and deep Q-learning under different vaccine costs. Results show that deep Q-learning with various combinations of architectures and learning rates suffers from convergence difficulties and stability issues. This highlights the practical advantage of tabular Q-learning over deep Q-learning in this application. Another alternative is the Dyna-Q algorithm, which augments the standard Q-learning algorithm by additional planning steps (Sutton and Barto, 2018). We also apply the Dyna-Q algorithm, and its performance is comparable with standard Q-learning in our application. Supplementary Materials S5 includes additional results on the comparisons between Dyna-Q and standard tabular Q-learning.

3.2.2 Policy interpretation

In Section 3.2.1, we determine a reasonable vaccine cost is around 0.04 based on the mortality rates ratio after 30 days of a booster and a severe infection. Based on the 20 replicates, we obtain a confidence measure for receiving booster of the Q-table-based policy. We determine a group of people need to receive the booster if more than 10 replicates out of the 20 suggest so (i.e., confidence measure for receiving booster is bigger than $10/20 = 0.5$).

When vaccine cost is set to 0.05, the Q-table policy recommends that adults aged 50-

65 without baseline immunosuppressant use delay the booster until the 7th month after their second vaccination. In contrast, individuals of the same age group with baseline immunosuppressant use, as well as adults older than 65 without immunosuppressant use, are advised to receive the booster earlier, at 5-6 months. Among those aged 30-50, the policy suggests receiving the booster at 5-6 months if there is no immunosuppressant use, but postponing until the 7th month if there is baseline immunosuppressant use. When the vaccine cost rises to 0.1, the recommended policies remain unchanged, although the associated confidence measures for booster administration generally decrease. Conversely, when the vaccine cost decreases to 0.04, all adult age groups, regardless of immunosuppressant status, are advised to receive the booster 5-6 months after the second vaccination. Further decreasing the cost to 0.03 does not alter the recommended timing, but confidence measures increase to at least 0.9 for most groups. Across costs around 0.04, the confidence in recommending a booster at 5-6 months is generally higher for individuals with baseline immunosuppressant use compared to those without, and higher for adults over 65 compared to younger adults. Overall, these results suggest prioritization of booster vaccination for older adults and those with immunosuppressant use which aligns with prior findings of previous COVID-19 vaccine studies (Risk et al., 2022a,b; Shen et al., 2022).

4 Comparisons with Deep-Q Learning

In this section, we compare tabular Q-learning and deep-Q learning in the booster policy learning application, using the same environment simulated through microsimulation Section 3.1. Unlike tabular Q-learning where the Q-function is explicitly stored in a table, deep-Q learning approximates the Q-function using a deep neural network parameterized by θ, i.e, $Q_\theta(s, a)$.

The deep Q-learning algorithm minimizes the temporal difference (TD) loss function, defined as $L(\theta) = \frac{1}{2}\{y - Q_\theta(s, a)\}^2$, where the target value $y = r(s, a) + \gamma \max_u Q_{\theta^-}(s', u)$

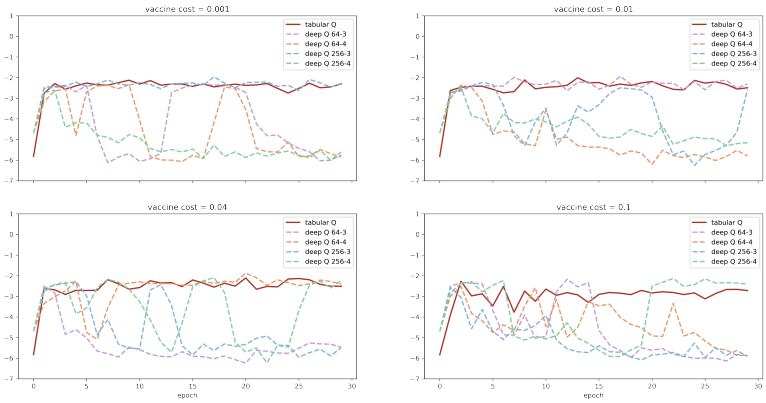


Figure 4: Comparison between tabular Q-learning and deep Q-learning with different architectures: the average reward evaluated over individuals along 30 epochs for different vaccine costs.

with $Q_{\theta^-}(s', u)$ representing the target network, which is periodically updated with the lagged parameters θ^- to stabilize training (Mnih et al., 2015). The parameters of the neural network are updated using gradient descent, following the rule $\theta \leftarrow \theta - \beta \nabla_\theta L(\theta)$ where $\nabla_\theta L(\theta) = -\delta \nabla_\theta Q_\theta(s, a)$ with approximated TD error $\delta = r(s, a) + \gamma \max_u Q_{\theta^-}(s', u) - Q_\theta(s, a)$. The gradient descent update rule can be written as $\theta \leftarrow \theta + \beta \delta \nabla_\theta Q_\theta(s, a)$, which mirrors the tabular Q-learning update rule in (1), i.e., $Q(s, a) \leftarrow Q(s, a) + \beta \delta^*$, where $\delta^* = r(s, a) + \gamma \max_u Q(s', u) - Q(s, a)$ represents the TD error. The TD error δ in deep Q-learning is an approximation of the true TD error, as it relies on the maximum Q-values calculated using the lagged parameters of the target network. Consequently, while tabular Q-learning offers convergence guarantees under standard conditions, deep Q-learning does not share the same theoretical guarantees of optimality due to the inherent approximation and instability introduced by function approximation.

In Figure 4, we present the average reward over individuals along 30 epochs for deep Q-learning, considering various neural network architectures and learning rates. For comparison, we also show the average reward for tabular Q-learning over the same number of training epochs. The deep Q-learning models use networks with two hidden layers, each

containing either 64 or 256 nodes. For each architecture, we consider two learning rates, 10^{-3} and 10^{-4}, with the Adam optimizer (Kingma and Ba, 2014). For instance, in Figure 4, the line with label "deep Q 64-4" corresponds to a Q-network of two hidden layers with 64 nodes each, trained by the Adam optimizer with learning rates 10^{-4}.

Figure 4 suggests that deep Q-learning suffers from convergence difficulties and stability issues for all combinations of architectures and learning rates. The Q-network with two layers of 256 nodes and trained with learning rate 10^{-3} (deep Q 256-3) converges when vaccine cost is 0.001, but does not converge for other vaccine costs. Similarly, the Q-network with two layers of 64 nodes, trained with learning rate 10^{-3} (deep Q 64-3), converges only when vaccine cost is 0.01. The Q-network with two layers of 64 nodes and trained with learning rate 10^{-4} (deep Q 64-4) converges only when vaccine cost is 0.04. The same Q-network architecture trained with the same learning rate exhibits highly variable performance depending on the vaccine cost. Even when deep Q-learning converges, its has similar reward compared to that of tabular Q-learning. In contrast, tabular Q-learning demonstrates consistent convergence and robust performance across all scenarios. This highlights the practical advantage of tabular Q-learning over deep Q-learning in public health applications, where the state and action spaces are often discrete and relatively small. Given the challenges associated with deep Q-learning, especially in terms of stability and convergence, tabular Q-learning remains a reliable and efficient choice when the problem structure allows for it.

5 Discussion and Conclusion

In this paper, we propose a novel framework combining tabular Q-learning with an RNN-based environment simulator to optimize COVID-19 booster vaccination policies. The proposed approach addresses key challenges in vaccine policy development, including the limitations of clinical trials and ethical concerns on need of direct interactions of the

Reinforcement Learning (RL) algorithms with the real world. By utilizing an RNN, we successfully create a digital twin of the infection dynamics for the target population, which models the temporal relationships of COVID-19 infections and vaccination status, generating simulated data that reflects real-world dynamics. The policy learned through our method outperforms the currently observed practices of COVID-19 booster vaccination, indicating its potential to enhance vaccine deployment and reduce infection rates.

A valid application of our framework relies on the Markov assumption. The RNN used in the microsimulation is capable of modeling both Markovian and non-Markovian transitions. We test whether the transition dynamics induced by the RNN satisfy the Markov property (Shi et al., 2020). Applied to the simulated data, this test indicates no evidence against the Markov assumption, which justifies the use of tabular Q-learning in our application. However, this justification is not at odds with the use of an RNN. While the RNN is capable of encoding non-Markovian dependence, it does not impose such a structure if the data do not support it. In applications where the Markov assumption is violated, one can mitigate non-Markovianity by augmenting the state with relevant past observation–action pairs (with the history length potentially determined adaptively, e.g., by Shi et al. (2020)) or by incorporating learned RNN representations, such as the final hidden layer, into the state definition. Such augmentations, however, would result in a state space that mixes discrete and continuous components, in which case a straightforward policy table is no longer directly available. Nonetheless, interpretable policy summaries can still be obtained for clinically relevant groups by marginalizing over the continuous state variables.

Our framework offers several advantages. First, the RNN-generated simulated data enables continuous exploration of potential policies without ethical concerns. This allows us to conduct extensive policy evaluations without requiring real-world interventions, avoiding the risks of harmful or suboptimal decisions. Second, by employing tabular Q-learning, we provide an interpretable and clear policy table, allowing policymakers to easily understand

and implement optimal vaccination strategies. While Deep Q-learning has been widely applied in healthcare for its flexibility in large and continuous state spaces, it suffers from convergence difficulties and stability issues in our applications. This instability highlights the value of tabular Q-learning, which, while simpler, offers more reliable and interpretable outcomes for public health problems where states and actions are discrete.

This research demonstrates the effectiveness of RL in public health policy development and presents a scalable solution for future pandemics or vaccine rollouts. Although we focus on COVID-19 booster policymaking as a case study, our approach is broadly applicable to problems requiring decision-making for different populations. Beyond vaccines for future pandemics and other public health interventions (e.g., smoking cessation strategies), a further example is cancer screening, where recommendations vary according to factors such as individual risk status, age, and family history (ACS, 2023).

Acknowledgement The authors thank the University of Michigan Data Office for assistance with data extraction from electronic medical records. Dr. Zhao's research is supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI158543. Dr. Kang's research was partially supported by National Institute of Health grants R01DA048993 and R01MH105561 and the National Science Foundation grant IIS-2123777. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

ACS (2023). American cancer society guidelines for the early detection of cancer.

Adebamowo, C., Bah-Sow, O., Binka, F., Bruzzone, R., Caplan, A., Delfraissy, J.-F., Heymann, D., Horby, P., Kaleebu, P., Tamfum, J.-J. M., et al. (2014). Randomised controlled trials for ebola: practical and ethical issues. *The Lancet*, 384(9952):1423–1424.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., et al. (2021). Efficacy and safety of the mrna-1273 sars-cov-2 vaccine. *New England Journal of Medicine*, 384(5):403–416.

CDC (2024). CDC museum COVID-19 timeline. `https://www.cdc.gov/museum/timeline/covid19.html`.

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, 40(5):373–383.

Chernozhukov, V., Kasahara, H., and Schrimpf, P. (2020). Causal impact of masks, policies, behavior on early covid-19 pandemic in the us. *Journal of Econometrics*, 220(1):23.

Eftekhari, H., Mukherjee, D., Banerjee, M., and Ritov, Y. (2020). Markovian and non-markovian processes with active decision making strategies for addressing the covid-19 pandemic. *arXiv preprint arXiv:2008.00375*.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR.

Gasparini, A. (2018). comorbidity: An r package for computing comorbidity scores. *Journal of Open Source Software*, 3(23):648.

Guo, H., Li, J., Liu, H., and He, J. (2022). Learning dynamic treatment strategies for coronary heart diseases by artificial intelligence: real-world data-driven study. *BMC Medical Informatics and Decision Making*, 22(1):39.

Julien, J., Ayer, T., Tapper, E. B., Barbosa, C., Dowd, W. N., and Chhatwal, J. (2022). Effect of increased alcohol consumption during covid-19 pandemic on alcohol-associated liver disease: a modeling study. *Hepatology*, 75(6):1480–1490.

Jüni, P., Altman, D. G., and Egger, M. (2001). Assessing the quality of controlled clinical trials. *BMJ*, 323(7303):42–46.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., Núñez, R. C., Cohen, J. A., Selvaraj, P., Hagedorn, B., et al. (2021). Covasim: an agent-based model of covid-19 dynamics and interventions. *PLOS Computational Biology*, 17(7):e1009149.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kompella, V., Capobianco, R., Jong, S., Browne, J., Fox, S., Meyers, L., Wurman, P., and Stone, P. (2020). Reinforcement learning for optimization of covid-19 mitigation policies. *arXiv preprint arXiv:2010.10560*.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014). Interactive model building for q-learning. *Biometrika*, 101(4):831–847.

Lander, J., Langhof, H., and Dierks, M.-L. (2019). Involving patients and the public in medical and health care research studies: An exploratory survey on participant recruiting and representativeness from the perspective of study authors. *PloS One*, 14(1):e0204187.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J., and Wang, Y. (2017). Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 380–385. IEEE.

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. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Monrad, J. T. (2020). Ethical considerations for epidemic vaccine trials. *Journal of Medical Ethics*, 46(7):465–469.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., et al. (2020). Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615.

Risk, M., Hayek, S. S., Schiopu, E., Yuan, L., Shen, C., Shi, X., Freed, G., and Zhao, L. (2022a). Covid-19 vaccine effectiveness against omicron (b. 1.1. 529) variant infection and hospitalisation in patients taking immunosuppressive medications: a retrospective cohort study. *The Lancet Rheumatology*, 4(11):e775–e784.

Risk, M., Shen, C., Hayek, S. S., Holevinski, L., Schiopu, E., Freed, G., Akin, C., and Zhao, L. (2022b). Comparative effectiveness of coronavirus disease 2019 (covid-19) vaccines against the delta variant. *Clinical Infectious Diseases*, 75(1):e623–e629.

Rutter, C. M., Zaslavsky, A. M., and Feuer, E. J. (2011). Dynamic microsimulation models for health outcomes: a review. *Medical Decision Making*, 31(1):10–18.

Shen, C., Lin, M., Lee, Y., Dong, M., and Zhao, L. (2024). State-of-the-art learning covid-19 vaccine effectiveness using lstm. *Informatics in Medicine Unlocked*, page 101561.

Shen, C., Risk, M., Schiopu, E., Hayek, S. S., Xie, T., Holevinski, L., Akin, C., Freed, G., and Zhao, L. (2022). Efficacy of covid-19 vaccines in patients taking immunosuppressants. *Annals of the Rheumatic Diseases*, 81(6):875–880.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. (2020). Does the markov decision process fit the data: Testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pages 8807–8817. PMLR.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of Machine Learning Research*, 15(1):1929–1958.

Sun, Q., Jankovic, M. V., Budzinski, J., Moore, B., Diem, P., Stettler, C., and Mougiakakou, S. G. (2018). A dual mode adaptive basal-bolus advisor based on reinforcement learning. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2633–2641.

Sutton, R. S. (1997). On the significance of markov decision processes. In *Artificial Neural Networks—ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings 7*, pages 273–282. Springer.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tan, J., Ge, Y., Martinez, L., Sun, J., Li, C., Westbrook, A., Chen, E., Pan, J., Li, Y., Cheng, W., et al. (2022). Transmission roles of symptomatic and asymptomatic covid-19 cases: a modelling study. *Epidemiology & Infection*, 150:e171.

Tian, T., Tan, J., Luo, W., Jiang, Y., Chen, M., Yang, S., Wen, C., Pan, W., and Wang, X. (2021). The effects of stringent and mild interventions for coronavirus pandemic. *Journal of the American Statistical Association*, 116(534):481–491.

Tseng, H.-H., Luo, Y., Cui, S., Chien, J.-T., Ten Haken, R. K., and Naqa, I. E. (2017). Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical Physics*, 44(12):6690–6705.

Van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Wan, R., Zhang, X., and Song, R. (2021). Multi-objective model-based reinforcement learning for infectious disease control. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1634–1644.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.

Weltz, J., Volfovsky, A., and Laber, E. B. (2022). Reinforcement learning methods in public health. *Clinical Therapeutics*, 44(1):139–154.

Wu, X., Li, R., He, Z., Yu, T., and Cheng, C. (2023). A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *NPJ Digital Medicine*, 6(1):15.

Yu, C., Dong, Y., Liu, J., and Ren, G. (2019). Incorporating causal factors into reinforcement learning for dynamic treatment regimes in HIV. *BMC medical informatics and decision making*, 19:19–29.

Yu, C., Liu, J., Nemati, S., and Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

Zhang, J. (2020). Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pages 11012–11022. PMLR.