EpidRLearn: Learning Intervention Strategies for Epidemics with Reinforcement Learning

Maria Bampa[1]✉, Tobias Fasth[1,2], Sindri Magnusson[1], and Panagiotis Papapetrou[1]

[1] Dept. of Computer and Systems Sciences, Stockholm University
[2] The Public Health Agency of Sweden
{maria.bampa, sindri.magnusson, panagiotis}@dsv.su.se
tobias.fasth@folkhalsomyndigheten.se

Abstract. Epidemics of infectious diseases can pose a serious threat to public health and the global economy. Despite scientific advances, containment and mitigation of infectious diseases remain a challenging task. In this paper, we investigate the potential of reinforcement learning as a decision making tool for epidemic control by constructing a deep Reinforcement Learning simulator, called `EpidRLearn`, composed of a contact-based, age-structured extension of the SEIR compartmental model, referred to as `C-SEIR`. We evaluate `EpidRLearn` by comparing the learned policies to two deterministic policy baselines.We further assess our reward function by integrating an alternative reward into our deep RL model. The experimental evaluation indicates that deep reinforcement learning has the potential of learning useful policies under complex epidemiological models and large state spaces for the mitigation of infectious diseases, with a focus on COVID-19.

Keywords: reinforcement learning · mitigation policies · COVID-19.

1 Introduction

The recent outbreak of the COVID-19 pandemic underlines the potentially catastrophic consequences of an epidemic outbreak not only for public health but also for the global economy and society as a whole. Moreover, identifying the most appropriate sequence of mitigation policies is a challenge [13], with the vast majority of the countries worldwide rapidly adopting mitigation measures to limit the impact of an ongoing pandemic. When making such decisions and studying the effect of prevention strategies on the population dynamics, officials usually rely on epidemiological models that predict and project the course of the epidemic. A classic example of such an epidemiological model is SEIR (see, e.g., Allen et al. [1]); the response to an emerging virus has been highly relying on such models for many decades, rendering them an effective tool for modeling, forecasting, and studying dynamics [4]. In the context of COVID-19, several studies have tried to assess the pandemic patterns using classic compartmental models while considering Non-Pharmaceutical Interventions (NPIs) [15].

Some other studies adopted extended versions of typical compartmental models to estimate epidemiological parameters and extrapolate the disease dynamics in the context of social distancing [3, 10]. However, when results are dependent on heterogeneous individuals' characteristics, these compartmental models are inadequate as they fail to capture such characteristics. On the other hand, the development of prevention strategies that fulfill various criteria remains challenging and can lead to a complex sequential decision-making problem. A solution would be to use agent-based models that explicitly track the current epidemic state of individuals while dynamically introducing restrictions and modeling the effect of government restrictions on the spread of the epidemic. Modern Reinforcement Learning (RL) algorithms are well-suited for the problem of optimizing government response to epidemics. RL problems are closed-loop problems, where the learning system's actions influence its later inputs; the agents in RL learn what to do and map situations to actions to maximize a numerical reward signal [14].

Related work. A wide body of research focused on creating optimal NPI mitigation policies with RL in the context of influenza and, more recently COVID-19 [6–8, 12], while a few other works also included vaccination policies [17]. In the context of COVID-19, Ohi et al. [12] developed a virtual environment that mimics the simple SEIR to account for the lack of randomness inherent in the epidemiological equations. The reward is based on a random number generated by the healthy population and on the percentage of active cases, while their actions are introduced as three levels of contact restrictions of the population. Moreover, Libin et al. [8] constructed an age-structured model and tried to optimize the opening and closing of schools in between regions in the United Kingdom in the context of influenza. Nonetheless, creating optimal mitigation policies in the context of an epidemic is not a trivial task and entails sufficient reasoning on the chosen epidemic models and the underlying optimization components. Providing policy-makers with reasons to trust these AI-based decision-making models requires: (1) an epidemic model that captures the disease dynamics while having the flexibility to fit real epidemic data, (2) an action space that can be generalized or specific enough according to the stakeholders' needs, and (3) a justification of the optimization component (reward function) that actively influences the decisions of the RL agent. Hence, there is a need for a general-purpose RL-based epidemic model that addresses the above requirements.

Contributions. (1) We present and employ an extended SEIR compartmental model, called `C-SEIR`, a six-compartment, age-structured, contact-reduction based model that additionally considers groups of people that are infected but unreported, and recovered with PCR-negative and positive tests; in that way, `C-SEIR` models a more realistic COVID-19 case; (2) We introduce `EpidRLearn`, a general-purpose deep RL simulator that models an epidemic, using `C-SEIR`, and entails a set of actions based on various population contact reduction levels. It further employs a reward function that captures the trade-off between the increase of infections and the ramifications of contact reduction measures on the population; (3) We provide an extensive evaluation of the strategies proposed by `EpidRLearn` and investigate the agent's proposed actions by considering baseline policies and epidemic data, and an assessment of the reward function by incorporating in our model an alternative reward function from a recent benchmark.

2 `EpidRLearn`: Intervention Strategies for Epidemics with Reinforcement Learning

We define the RL problem by elaborating on its four components: environment, observations, actions, and reward.

The `C-SEIR` environment. The RL environment of `EpidRLearn` is based on `C-SEIR`, the contact-based, age-structured SEIR compartmental model proposed in this paper. Our model extends the structure of the earlier SEIR models (susceptible (S), exposed (E), infected (I), and recovered (R), see, e.g., [8, 12]) by (i) dividing compartment I into I^r (reported cases) and I^u (unreported cases), and (ii) dividing the R compartment into R^+ (PCR-positive population) and R^- (PCR-negative population). `C-SEIR` is defined as a system of Ordinary Differential Equations per age group $G_i \in [1, q]$:

$$\begin{aligned}
\frac{dS_i}{dt} &= -\beta_i S_i \sum_{j\in[1,g]} C_{ij} P_{ij} (\alpha I^u_j + I^r_j)/N_j \\
\frac{dE_i}{dt} &= \beta_i S_i \sum_{j\in[1,g]} C_{ij} P_{ij} (\alpha I^u_j + I^r_j)/N_j - \mu E_i \\
\frac{d{I^u}_i}{dt} &= \mu\eta E_i - \gamma {I^u}_i \\
\frac{d{I^r}_i}{dt} &= \mu(1-\eta) E_i - \gamma {I^r}_i \\
\frac{dR^+_i}{dt} &= \gamma I^r_i + \gamma I^u_i - \psi R^+_i \\
\frac{dR^-_i}{dt} &= \psi R^+_i .
\end{aligned} \tag{1}$$

`C-SEIR` is designed to simulate the spread of COVID-19 within and between a set of q population age groups $\mathbb{G} = \{G_1, \ldots, G_q\}$. A feature of `C-SEIR` is the usage of a $q \times q$ contact matrix C and a $q \times q$ contact reduction scalar matrix P. Each C_{ij} corresponds to the average contact frequency of an individual in age group G_i with an individual in age group G_j, while each P_{ij} is the contact reduction scalar value applied to C_{ij}. The risk of an individual in G_i getting infected by an individual in G_j is based on the transmission risk β_i, with

$$\beta_i = C_{ij} P_{ij} * (\alpha {I^u}_j + {I^r}_j)/N_j \ , \tag{2}$$

where α is an infectivity reduction scalar. The main role of P is to capture how the populations' contact patterns have changed throughout the pandemic. More concretely, for each day $t \in \{0, .., 365\}$ in the simulation the contact matrix is scaled by $P_{ij} \in [0, 1]$. A scalar of 0 implies no contacts between individuals in the corresponding age groups, while a scalar of 1 implies pre-pandemic contacts.

After getting exposed, the population stays in E for $\mu^{-1} = 5.1$ days (exposed rate) [9] on average before moving to either compartment I^r or I^u. Parameter η in compartments I^r or I^u defines the share of unreported cases. In these compartments, the population is infectious for $\gamma^{-1} = 5$ days (recovery rate) [16] and can during that time transmit the virus to the S population. The infected population moves to the R^+ compartment where it remains for an additional $\psi^{-1} = 5$ days (recovery rate). The population in the infected compartments (I^r, I^u) and the first recovered compartment (R^+) is assumed to be PCR-positive, i.e., the total time of PCR-positivity is ten days [5].

To demonstrate the utility of C-SEIR in our empirical evaluation (Sec. 4), we instantiate it to support three age groups ($q = 3$), i.e., G_1: 0-19, G_2: 20-69, and G_3: 70+ years old; but without loss of generality, our formulation holds for any number of age groups q and any configuration of age ranges. As a result, C becomes a 3×3 matrix, the configuration of which is based on an earlier epidemic study [11] and adapted to the Swedish demographics. In order to account for randomness in C-SEIR and the predictions, we transform the system of Ordinary Differential Equations (ODEs) into Stochastic Differential Equations (SDEs) by adding a Wiener process to each transition in the ODEs, hence adding stochastic noise [2]. The SDEs are evaluated at discrete time steps using the Euler-Maruyama approximation method [2].

Observations. The environment is observed after the agent has taken some action (at each time point t, i.e., each week). These observations S are passed to the agent serving as estimable information that can define the next action to be taken. This results in a total of $6 \times q$ parameters, one for each of the six compartments and one for each of the q age groups.

Actions. In order to reduce the spread of the epidemic, we need to limit the frequency of contacts between population groups. We hence define a set of actions $\mathcal{A}$ that impose population movement reductions. For simplicity and without loss of generality, we let the set of actions $\mathcal{A}$ correspond to four population movement reduction levels, i.e., $\mathcal{A} = \{A_0, A_1, A_2, A_3\}$, with A_0 describing a freely moving population with no imposed restrictions (Level 0), A_1 a movement reduction by 25% (Level 1), A_2 a movement reduction by 50% (Level 2), A_3 a movement reduction by 75% (Level 3). We note that the chosen action space is an estimation of the level of contact reductions in the population. While alternative (and additional) actions can be defined, an extensive definition of policies is not the main goal of this paper. Using any set of actions $\mathcal{A}$, we calibrate the movement of the population by modifying the contact reduction scalar P. As this scalar matrix can change per time unit t, we denote its configuration at time t as $P^{(t)}$. Before the start of an infectious disease, $P^{(0)}$ is an all-ones (i.e., unit) matrix, since the population still moves freely (following C). As the epidemic evolves and mitigation policies are imposed by $\mathcal{A}$, P is adjusted accordingly to reflect the movement reduction rate. At each time point t, the initial contact matrix C is scaled by P^t yielding a modified contact matrix $C^t = P^t \cdot C$ that alters the population contact pattern. Without loss of generality, for our experimental setup, we assume the following configuration for P^t given action $a \in \mathcal{A}$:

$$P^t(a) = \begin{cases} P^{(0)}, & \text{if } a = A_0 \textit{ (Level 0)} \\ 0.75 \times P^{(0)}, & \text{if } a = A_1 \textit{ (Level 1)} \\ 0.50 \times P^{(0)}, & \text{if } a = A_2 \textit{ (Level 2)} \\ 0.25 \times P^{(0)}, & \text{if } a = A_3 \textit{ (Level 3)} \end{cases} \tag{3}$$

Reward. Our reward function considers the trade-off between keeping the population contacts as unrestricted as possible while maintaining the infectious population low. The rationale is that we want to minimize the number of infected people while focusing on the populations' well-being; imposing interventions in the form of contact reductions in the population can have a negative impact on socioeconomic factors. The reward function is, hence, defined as follows:

$$r(s,a) = W_1 \times \texttt{FreeToMove}(s,a) \texttt{ + } W_2 \times \texttt{InfectionState}(s,a) \ . \tag{4}$$

The fraction of healthy population at state s is $H(s) = \frac{S+E+R^-}{\texttt{TotalPopulation}}$. Then the first component of the reward can be defined as a function of $H(s)$ and each action $a \in \mathcal{A}$. Following our earlier instantiation of $\mathcal{A}$, the first component of the reward, denoted as $\texttt{FreeToMove}(\cdot)$, is defined as

$$\texttt{FreeToMove}(s,a) = \begin{cases} H(s), & \text{if } a = A_0 \\ 0.75 \times H(s), & \text{if } a = A_1 \\ 0.5 \times H(s), & \text{if } a = A_2 \\ 0.25 \times H(s), & \text{if } a = A_3 \end{cases} \tag{5}$$

The second component of the reward, denoted as $\texttt{InfectionState}(\cdot)$, reflects the degree to which the infection decreases after an action is taken, i.e.,

$$\texttt{InfectionState}(s,s',a) = \sum_{i=1}^{q} \texttt{IS}_i(s,s',a) \ , \tag{6}$$

$\texttt{IS}_i(s,s',a)$ is the infection rate indicator for population group i, such that

$$\texttt{IS}_i(s,s',a) = \begin{cases} +0.5, & \text{if } I_i(s') < I_i(s) \ \& \ a = A_0 \\ -0.5, & \text{if } I_i(s') \geq I_i(s) \ \& \ a = A_0 \\ +0.5, & \text{if } I_i(s') > I_i(s) \ \& \ a \in \{A_1, A_2, A_3\} \\ -0.5, & \text{if } I_i(s') \leq I_i(s) \ \& \ a \in \{A_1, A_2, A_3\} \end{cases} \tag{7}$$

Training EpidRLearn. We employ the Proximal Policy Optimization (PPO) algorithm to train our agent on data during the Spring/Summer period (period 1) and evaluate the policies using data from Autumn/Winter (period 2). The agent is trained for 10000 episodes with $\gamma = 0.99$, and at the end of each episode the environment is reset to an initial state. The policy is evaluated over 10 runs. The network consists of 1 hidden layer of 128 units with the Tahn activation function. Training is episodic (25 simulated weeks per episode). The benchmark policies are also considered during Autumn/Winter, starting after week 25. The reader may refer to the supplementary material and source code found in the Github repository [3] for additional information on model parameters and verification.

3 Model Instantiation

Fitting C-SEIR to Real Data. We demonstrate the steps for calibrating the C-SEIR parameters by fitting it to real-world data. We choose a region in Sweden as our COVID-19 use-case since Sweden is one of the few countries that followed

[3] https://github.com/mariabampaai/epidrlearn

a set of nonrestrictive recommendations. Considering an epidemic of two waves, the model fitting is divided into two periods, period 1 (Feb. 27, 2020 - Aug. 23, 2020) and period 2 (Aug. 24, 2020 - Dec. 31, 2020), with transmission risks β_{p1} and β_{p2}, respectively. We also estimate the contact reduction scalar P, i.e., the daily relative contact change compared to before the pandemic.

In Fig. 4 (left sub-figure), we show the `C-SEIR` model projection in conjunction with the reported case data indicating that we can closely match the epidemic trends. The same figure also depicts the daily number of observed and fitted infected over period 1 and period 2. Note that our model reports the number of currently infected (prevalence) while the use-case data reports the number of newly infected people per day (incidence). Therefore, we define the incidence of reported cases for each age-group i from `C-SEIR` as follows:

$$\texttt{IncidenceInfected}_i = E_i/\mu \times (1-\eta) \tag{8}$$

Baseline Policies. To empirically evaluate the convergence of `EpidRLearn` to an optimal policy we need to devise baseline policies. The baseline policies will serve as an approximation to the actual policies followed by Sweden and will allow us to compare the policy created by `EpidRLearn` to the approximate true ones. The baseline policies are created by 'enforcing' the agent to choose specific actions on pre-defined steps/weeks of the episode.

Policy I (Sweden ECDC) The first policy corresponds to the mitigation measures announced by the authorities of Sweden as reported by ECDC [4]. The mitigation measures listed in ECDC entail a list of interventions followed by the Swedish authorities, among other countries. For a fair model comparison, we need to map these interventions to the corresponding `EpidRLearn` actions in Eq. 3. In that way, the mapping of the baseline policies to the ECDC response data is an abstraction of the actual policies followed by Sweden.

Specifically, the Swedish policy, denoted as π_{SE}, is defined as follows:

$$\pi_{SE} = \begin{cases} a = A_0, & \text{if } w_i < 6 \\ a = A_1, & \text{if } 6 \leq w_i \leq 22 \\ a = A_2, & \text{if } w_i > 22 \end{cases} \tag{9}$$

with w_i denoting the i^{th} week from the onset of the pandemic. In other words, we presume that Sweden placed first a series of soft restrictions on the 6^{th} week of the pandemic (Level 1 restrictions, A_1), which were further strengthened after the 22^{nd} week (Level 2 restrictions, i.e., A_2).

Policy II (Use-case Fitted) In addition, our intuition is that it is not only the policies of a governmental authority that affect the development of a pandemic but also the degree to which people comply with those policies and recommendations. Motivated by this, our second baseline policy corresponds to the fitted output obtained for the contact reduction scalar P_{ij} as derived in Sec. 3, i.e., the scalar that captures how the population's contact patterns have changed throughout the pandemic. This fitted output reflects the actual compliance to

[4] https://www.ecdc.europa.eu/en/publications-data/download-data-response-measures-covid-19

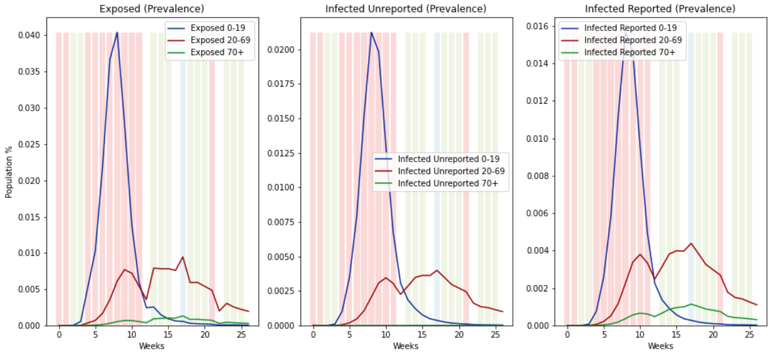

Fig. 1. The recommended optimal policy proposed by `EpidRLearn`. Red bars indicate Level 3, blue bars Level 2, green bars Level 1, and white bars Level 0 restrictions.

the chosen policy (in our case the one defined by the Swedish authorities). After the curve fitting step, P_{ij} depicts the reductions in C_{ij} (the contact matrix) as seen in the observed data. Hence when the fitted scalar P is applied to C_{ij} it reduces the contact pattern of the various groups as seen in the real COVID-19 data. This, we presume, abstractly depicts the interventions taken by the Swedish government in the form of contact restrictions. As the fitted P_{ij} already measures all the interventions, we enforce the RL agent to only take Level 0 actions (i.e., A_0), hence defining the policy as follows: $\pi_{SE-F} = A_0$, if $w_i >= 0$.

4 Empirical Evaluation

Optimal policy estimation and reward convergence. Fig. 1 depicts the resulting policy for period 2. From right to left, we see the Prevalence of Exposed, Infected Unreported, and Infected Reported per age group; each picture portrays the actions as colored bars (red for Level 3, blue for Level 2, green for Level 1, and white for Level 0 restrictions).We notice that in the first approximately 10 weeks, the policy primarily recommends Level 3 restrictions, potentially to minimize and slow down the peak of the infected population. Following that, the learned policy oscillates between actions within levels 0-3, and when the infection rate of age groups 20-69 and 70+ ascends, chooses one week of Level 2-3 restrictions. As the infectious population reduces, the agent mainly chooses actions of Level 0 to 1, mainly utilizing Level 1 restrictions.

We additionally compared the learned policy to the infection peak of the left sub-figure of Fig. 2 (the natural course of the epidemic as the sum over all age groups) and the baseline policies. The right sub-figure of Fig. 2 depicts the incidence curves of the learned policy over all age groups. We observe that `EpidRLearn` manages to reduce the incidence peak by at least a half compared to the epidemic's natural course. The reduction in the peak is expected as the agent enforces various levels of contact reductions.

Assessment of optimal policies. We demonstrate that `EpidRLearn` has learned something useful by comparing the incidence peaks and reward distributions with

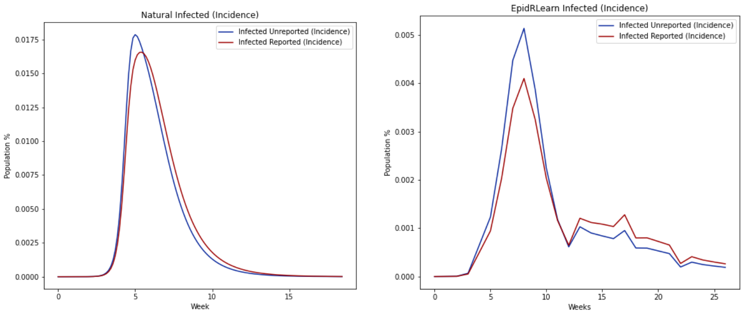

Fig. 2. Left: Incidence curves of the natural course of the epidemic Right: Incidence curves of the recommended policy. Whole population, reported and unreported cases.

the baseline policies. Our results in Fig. 3 indicate that the RL agent identifies a mitigation policy that optimizes the defined reward by reducing the pandemic's peak, ”flattening the curve” compared to the two baseline policies. More concretely, Fig. 3 compares the `EpidRLearn` policy to the baseline policies (Sweden and Sweden Fitted) for period 2. Comparing the incidence curves resulting from `EpidRLearn` policy to the Swedish baselines, we notice that the `EpidRLearn` policy reduces and shifts the incidence peak, i.e., “flatten the curve”.

However, the reward of the fitted Swedish policy (in yellow color) is considerably higher; potentially due to the design of this baseline, since the contact reduction scalar already measures all the interventions for the fitted use-case, the RL agent chooses only actions of Level 0 (i.e., A_0) directing the reward to higher values. More importantly, we should emphasize that the provided policies are ’theoretically’ optimal and coupled with the hypothetical use-case scenario studied here. In order to draw realistic conclusions for particular use-cases, e.g., Sweden, one would need to have access to the real and complete epidemiological data of each use-case, consider the possibility of new variants and re-calibrate `C-SEIR` to depict the previously mentioned. However, we can safely conclude that `EpidRLearn` can function as a strong assisting tool for the public health authorities to provide recommendations that reduce the epidemic spread by considering both the severity of the pandemic as well as other factors related to the general well-being of the population.

Assessment of Reward function. We finally provide an evaluation of the reward function by incorporating in our model an alternative function from Libin et al. [8], as we believe their environment implementation is closer to that of `EpidRLearn`; on the other hand the model that was proposed by [12] is based on a more granular level of SEIR and a virtual representation of the population, rendering the combination of their set-up and reward less suitable for our work. We compare the incidence curves in Fig. 4 (right) and notice that `EpidRLearn` reduces and shifts the epidemic peak compared to the alternative reward function. This demonstrates the flexibility of `EpidRLearn`'s reward component and the adjustability of `EpidRLearn` to incorporate in the model the stakeholders' requirements in the form of a reward function. For additional information on em-

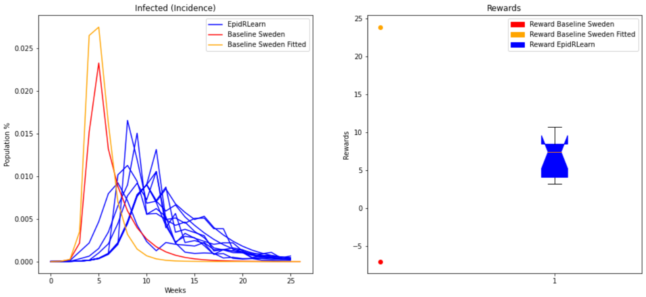

Fig. 3. Left: Comparison of `EpidRLearn` and Baselines, for each case reported is the sum of incidence infected (reported and unreported), Right: Rewards for `EpidRLearn` over 10 simulated runs and Baselines.

pirical verification, sensitivity analysis, and experiments, the reader may refer to the supplementary material in the provided GitHub repository.

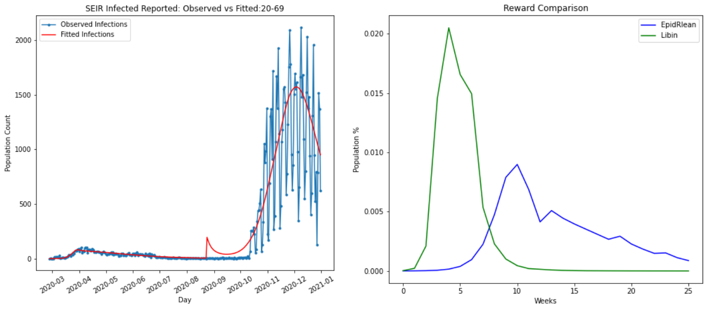

Fig. 4. Left:Fitting `C-SEIR` on the Swedish use-case data for age group 20-69. y-axis incidence of reported cases. Right: Comparison of the incidence infected (sum of reported and unreported) of the alternative to EpidRLearn reward. y-axis population %.

5 Conclusions

Finding an optimal policy to reduce the spread of an epidemic can be a challenging task in the space of unlimited interventions and potential socio-economic constraints. We demonstrated the potential of epidemic mitigation control using deep RL for COVID-19. The policy learned by `EpidRLearn` limits epidemic spread while taking into account the ramifications of continuous contact restrictions. Future work includes the study of a more granular epidemiological model, the consideration of mobility data, and contact restrictions between regions.

Acknowledgements. This work was supported by the Vinnova grant on “AI-supported design of more effective intervention strategies” and the Digital Futures EXTREMUM project.

References

1. Allen, E.J., Allen, L.J., Arciniega, A., Greenwood, P.E.: Construction of equivalent stochastic differential equation models. Stochastic analysis and applications **26**(2), 274–297 (2008)
2. Allen, L.J.: A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. Infectious Disease Modelling **2**(2), 128–142 (2017)
3. Cao, J., Jiang, X., Zhao, B., et al.: Mathematical modeling and epidemic prediction of covid-19 and its significance to epidemic prevention and control measures. Journal of Biomedical Research & Innovation **1**(1), 1–19 (2020)
4. Cobey, S.: Modeling infectious disease dynamics. Science **368**(6492), 713–714 (2020)
5. Hu, Z., Song, C., Xu, C., Jin, G., Chen, Y., Xu, X., Ma, H., Chen, W., Lin, Y., Zheng, Y., et al.: Clinical characteristics of 24 asymptomatic infections with covid-19 screened among close contacts in nanjing, china. Sc. China Life Sciences **63**(5), 706–711 (2020)
6. Khalilpourazari, S., Doulabi, H.H.: Designing a hybrid reinforcement learning based algorithm with application in prediction of the covid-19 pandemic in quebec. Annals of Oper. Research pp. 1–45 (2021)
7. Kwak, G.H., Ling, L., Hui, P.: Deep reinforcement learning approaches for global public health strategies for covid-19 pandemic. Plos one **16**(5) (2021)
8. Libin, P., Moonens, A., Verstraeten, T., Perez-Sanjines, F., Hens, N., Lemey, P., Nowé, A.: Deep reinforcement learning for large-scale epidemic control. Tech. rep. (2020)
9. Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.m., Yuan, B., Kinoshita, R., Nishiura, H.: Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. Journal of clinical medicine **9**(2), 538 (2020)
10. Liu, Z., Magal, P., Seydi, O., Webb, G.: A model to predict covid-19 epidemics with applications to south korea, italy, and spain. medRxiv (2020)
11. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., et al.: Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med **5**(3), e74 (2008)
12. Ohi, A.Q., Mridha, M., Monowar, M.M., Hamid, M.A.: Exploring optimal control of epidemic spread using reinforcement learning. Scientific reports **10**(1), 1–19 (2020)
13. Richard, Q., Alizon, S., Choisy, M., Sofonea, M.T., Djidjou-Demasse, R.: Age-structured non-pharmaceutical interventions for optimal control of covid-19 epidemic. PLOS Computational Biology **17**(3), 1–25 (03 2021)
14. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
15. Wang, C., Liu, L., Hao, X., Guo, H., Wang, Q., Huang, J., He, N., Yu, H., Lin, X., Pan, A., et al.: Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of coronavirus disease 2019 in wuhan, china. MedRxiv (2020)
16. Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., et al.: Virological assessment of hospitalized patients with covid-2019. Nature **581**(7809), 465–469 (2020)
17. Yaesoubi, R., Cohen, T.: Dynamic health policies for controlling the spread of emerging infections: influenza as an example. PloS one **6**(9), e24043 (2011)