Appendix

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

1 Methods

The RL environment is based on `C-SEIR` the contact-based, age-structured SEIR compartmental model proposed in this paper. Refer to Table 1 for an explanation of the `C-SEIR` parameters.

Table 1. The variables and parameter initialization of the proposed `C-SEIR` model. The second column describes each variable, while the third column provides the values of the three age groups used in our instantiation in the experimental evaluation. These values correspond to the values for the Swedish case.

	Description	**Value [0-19, 20-69, 70+]**
N	population	[500000, 1500000, 250000]
S	susceptible	[500000, 1499924, 250000]
E	exposed	[0, 0, 0]
I^r	infected (reported)	[0, 1, 0]
I^u	infected (unreported)	[0, 75, 0]
R^+	recovered (PCR-positive)	[0, 0, 0]
R^-	recovered (PCR-negative)	[0, 0, 0]
α	infectivity reduction	[0.5, 0, 0]
β_i	general transmission risk	
β_{p1}	transmission risk (period 1)	[0.235, 0.157, 0.260]
β_{p2}	transmission risk (period 2)	[0.446, 0.041, 0.054]
μ	exposed rate	1/5.1
γ	recovery rate	1/5.0
ψ	recovery pcr rate	1/5.0
C	contact matrix	
P_{ij}	contact reduction scalar	

A configuration of C for the Swedish case is provided in Table 2. Each cell in C_{ij} corresponds to the average contact frequency of an individual in age group G_i with an individual in age group G_j.

Table 2. The contacts within and between age groups 0-19, 20-69 and 70+.

	0-19	**20-69**	**70+**
0-19	7.452397574	4.718777779	0.290063626
20-69	1.764106486	8.544229624	0.624169322
70+	0.443861795	2.55482785	1.69

2 Verification on extreme cases.

We use two extreme cases to verify that `EpidRLearn` works as expected. In the first extreme case, we implement a reward function that assigns `FreeToMove` with a weight of zero, expecting to see a learned policy where the agent does not take into account the ramifications of restricting the movement of the population. In the second extreme case, we assign `InfectionState` with a weight of zero, allowing the system to choose actions that will maximize the number of people that can move without restrictions. The results of these extreme cases are presented in Fig. 1, where each colored vertical bar corresponds to a measure taken by `EpidRLearn` (a red bar for Level 3, a blue bar for Level 2, a green bar for Level 1 and a white bar for Level 0 restrictions). For $W_1 = 0$, $W_2 = 1$ (upper sub-figure), i.e., free movement is of no importance as long as the infection rate is minimized, the agent chooses to take various levels of contact reduction. For $W_1 = 1$, $W_2 = 0$ (lower sub-figure), i.e., free movement is of utmost importance while infection rate does not matter, the agent learns a policy that does not contain any level of contact reduction.

3 Sensitivity analysis of the rewards.

Continuous contact restrictions impose an extreme burden to the well-being of the population and to the economy, and hence they should be avoided. Following that, we conduct a sensitivity analysis with respect to the reward's weights to evaluate which combination of them yields the lowest epidemic incidence curve. By controlling the weights in the reward we can compare several curves with different importance given on the freely moving population and the minimization of infection. We consider four different combinations of weights and train and evaluate the resulting policies only during period 1 (Spring/Summer). We choose to proceed with the combination of weights that yields one of the lowest incidence peak, i.e., $W_1 = 0.4$, $W_2 = 0.6$. Note that the combination of weights depends on what the stakeholder wants to achieve and in that way can be changed. See Fig. 2 for a visual representation of the sensitivity analysis.

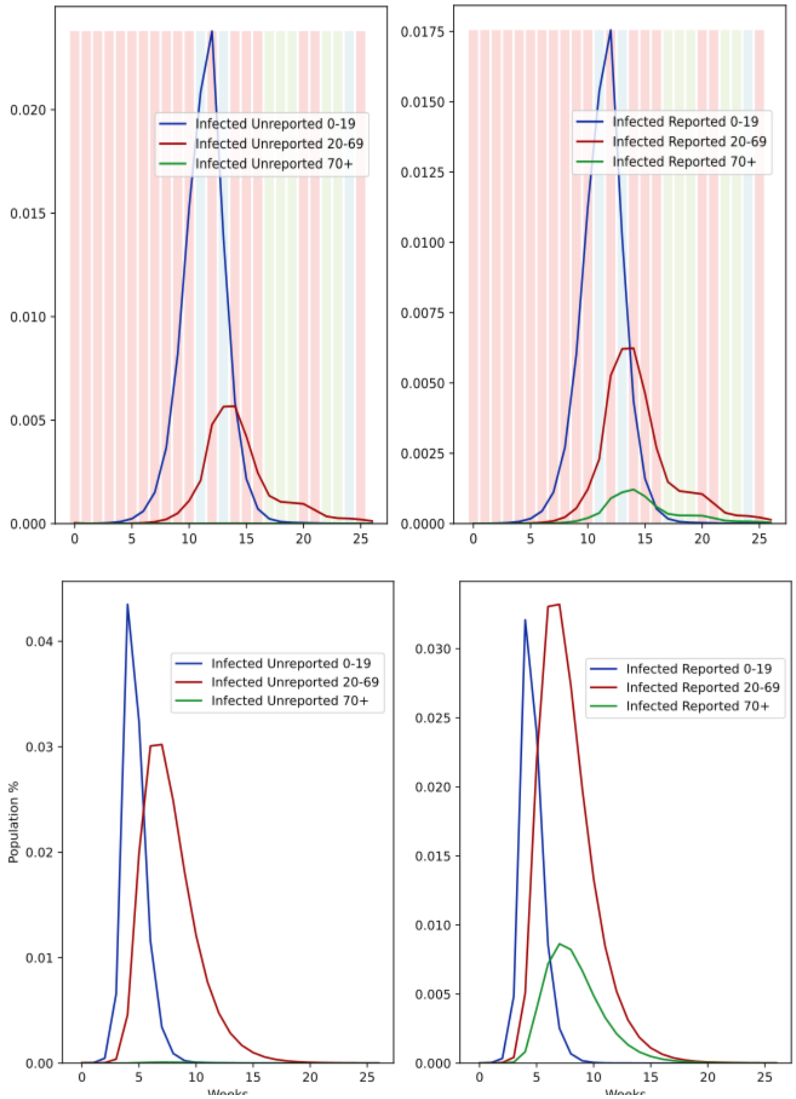

Fig. 1. Verification on two extreme cases. (Top) Red bars indicate level 3 restrictions, blue bars level 2 restrictions, green bars level 1 restrictions, and white bars level 0 restrictions; (Bottom) the corresponding incidence curves for unreported and reported cases per age group.

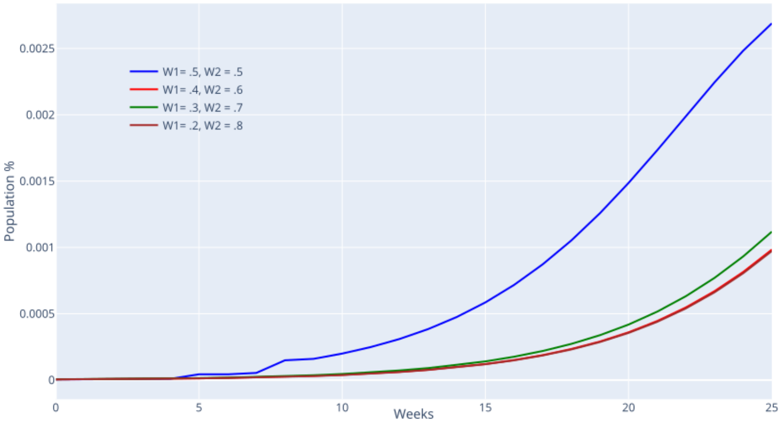

Fig. 2. Sensitivity analysis for the reward function.

4 Results

Fig. 4 depicts a single run of the ODEs with nor restrictions or the implementation of contact reduction scaling and Fig. 3 depicts the respective incidence curves of infected reported and unreported as sum over the age groups. Fig. 5 depicts the reward convergence demonstrating that `EpidRLearn` manages to learn an optimal policy as the reward is maximized (and converges). Figure 6 depicts the incidence curves of the learned policy per age-group (left: Unreported, right: Reported).

We evaluate `EpidRLearn` on period 2, using as initial population parameters the ones from the last week of period 1. See Fig. 7 for the optimal policy recommendation for this case.

5 Assessment of Reward Function

In Libin et al. [1], in order to reduce the attack rate, the authors measure the negative loss in the susceptible population over one simulated week and include a budget depletion for their school closure policy. To incorporate this reward function in our work we design a budget of 6 weeks for Level 3 and Level 2 movement restrictions, respectively, and measure the negative loss in the susceptible population as $r(s, \alpha) = -(S - S')$. Once the budget is depleted the agent is not able to choose Level 2 or 3 actions anymore. Fig. 8 depicts the learned policy using the previously mentioned reward function; we observe that in this case the agent enforces Level 2-3 movement restrictions after the peak of the 0-19 age group and near the peaks of the 20-69 and 70+ age groups, while for the rest of the weeks oscillates between actions of Level 0 to 1.

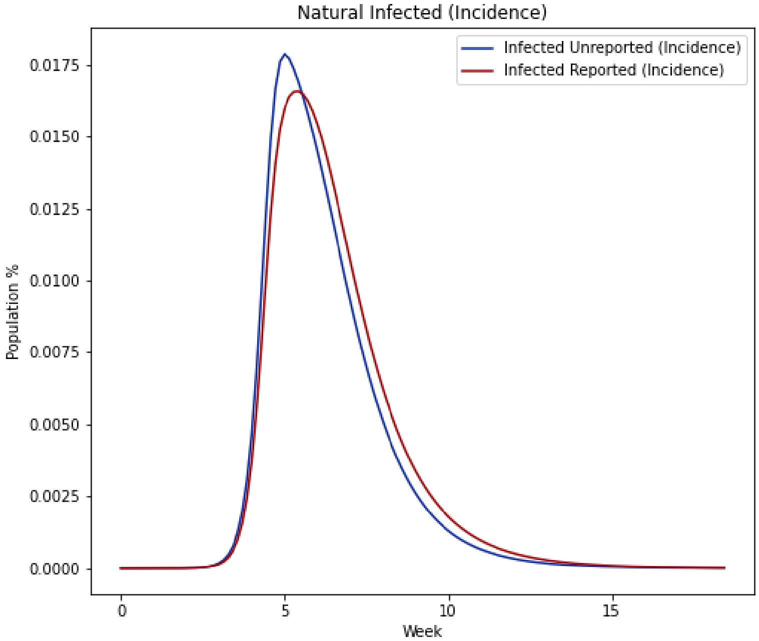

Fig. 3. Natural course of the epidemic of Infected Incidence, with no contact restrictions from `C-SEIR` for all age-groups. We observe both the reported and unreported cases as a fraction of the population.

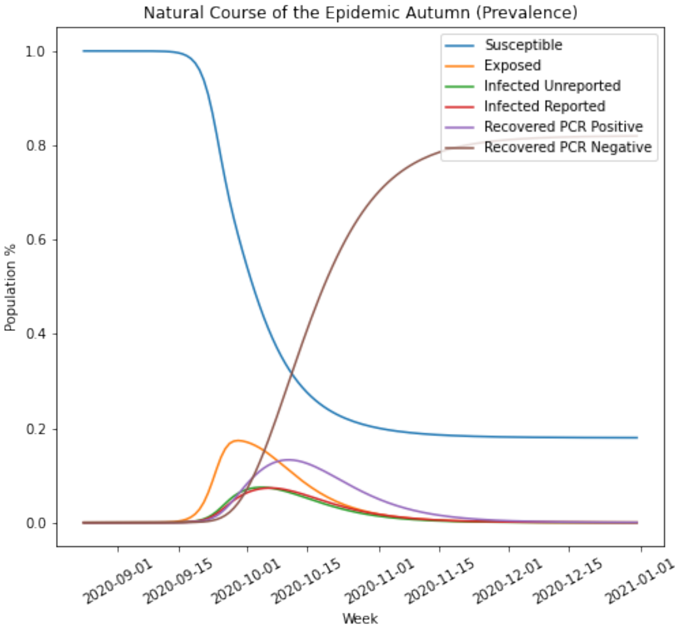

Fig. 4. Natural course of the epidemic, with no contact restrictions from `C-SEIR`, for all age-groups. The figure depicts how the six compartments of the model develop over time.

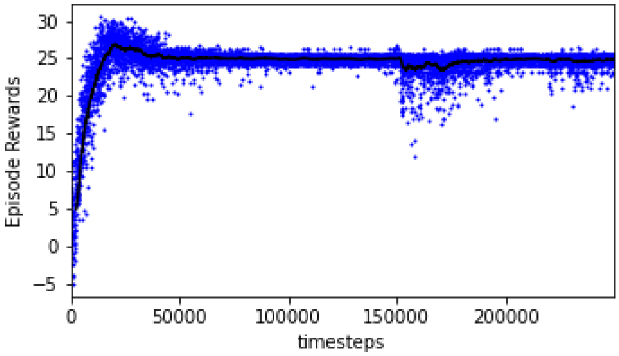

Fig. 5. Reward for each time step and reward rolling mean, of `EpidRLearn` for the chosen weight combination after 10000 episodes, indicating that the agent learns an optimal policy.

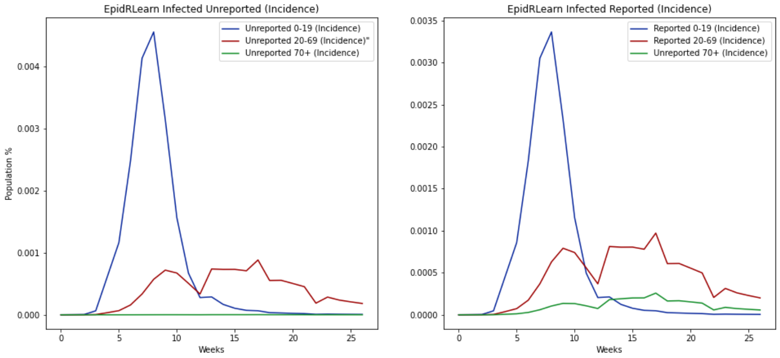

Fig. 6. Incidence curves of the recommended policy per age group for unreported (left-hand) and reported (right-hand) cases.

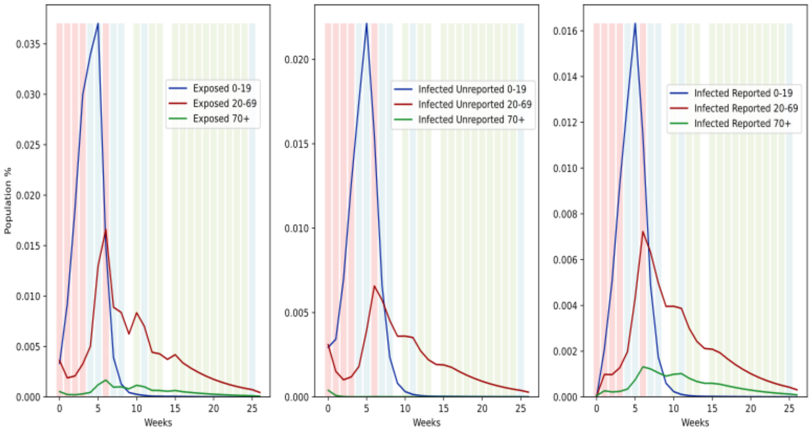

Fig. 7. The recommended optimal policy proposed by `EpidRLearn`. Red bars indicate level 3 restrictions, blue bars level 2 restrictions, green bars level 1, and white bars level 0 restrictions. The policy was trained with initial data for the population from period1 (spring).

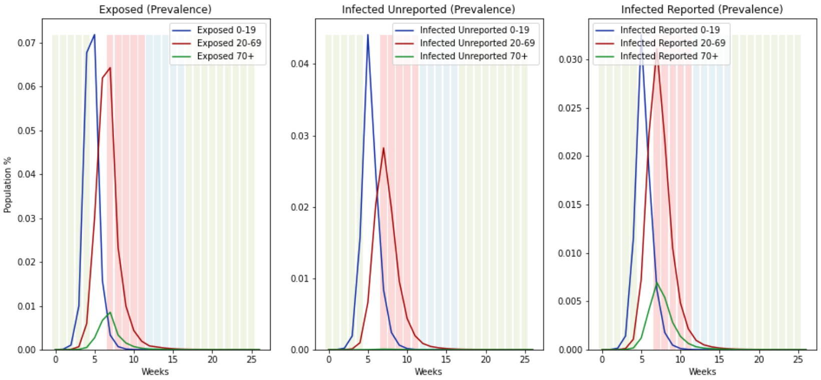

Fig. 8. The recommended optimal policy proposed using the alternative reward function. Red bars indicate level 3 restrictions, blue bars level 2 restrictions, green bars level 1, and white bars level 0 restrictions.

References

1. 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)