Efficient evaluation of influenza mitigation strategies using preventive bandits

Pieter Libin
Artificial Intelligence Lab
Vrije Universiteit Brussel
pieter.libin@vub.ac.be

Timothy Verstraeten
Artificial Intelligence Lab
Vrije Universiteit Brussel
tiverstr@vub.ac.be

Kristof Theys
Rega Institute
Katholieke Universiteit Leuven
kristof.theys@kuleuven.be

Diederik Roijers
Artificial Intelligence Lab
Vrije Universiteit Brussel
diederik.roijers@vub.ac.be

Peter Vrancx
Artificial Intelligence Lab
Vrije Universiteit Brussel
peter.vrancx@vub.ac.be

Ann Nowé
Artificial Intelligence Lab
Vrije Universiteit Brussel
ann.nowe@vub.ac.be

ABSTRACT

Pandemic influenza has the epidemiological potential to kill millions of people. While different preventive measures exist, it remains challenging to implement them in an effective and efficient way. To improve preventive strategies, it is necessary to thoroughly understand their impact on the complex dynamics of influenza epidemics. To this end, epidemiological models provide an essential tool to evaluate such strategies *in silico*. Epidemiological models are frequently used to assist the decision making concerning the mitigation of ongoing epidemics. Therefore, rapidly identifying the most promising preventive strategies is crucial to adequately inform public health officials. To this end, we formulate the evaluation of prevention strategies as a multi-armed bandit problem. The utility of this novel evaluation method is validated through experiments in the context of an individual-based influenza model.

We demonstrate that it is possible to identify the optimal strategy using only a limited number of model evaluations, even if there is a large number of preventive strategies to consider.

CCS Concepts

•**Theory of computation** → **Reinforcement learning;**
•**Computing methodologies** → **Multi-agent systems;**
•**Applied computing** → **Computational biology;**

Keywords

epidemiological models, preventive strategies, pandemic influenza, multi-armed bandits, reinforcement learning

1. INTRODUCTION

The influenza virus is responsible for the deaths of half a million people each year [38]. Additionally, seasonal influenza epidemics cause a significant economic burden [30]. While influenza is typically confined to local epidemics, it is possible for influenza to cause a pandemic when a novel strain emerges that spreads easily among human hosts worldwide [31]. Pandemic influenza occurs less frequently than seasonal influenza but the outcome with respect to morbidity and mortality can be much more severe, potentially killing millions of people worldwide [33]. Therefore, it is essential to study mitigation policies to control pandemic influenza epidemics.

For influenza, different preventive measures exist: i.a., vaccination, social measures (e.g., school closures) and antiviral drugs. However, the efficiency of these measures greatly depends on their availability, as well as on epidemiological characteristics. Furthermore, governments typically have limited resources to implement such measures. Therefore, it remains challenging to formulate prevention strategies that make effective and efficient use of these preventive measures while putting as little strain on the available resources as possible. To improve the development of preventive strategies, it is necessary to thoroughly understand the complex dynamics of influenza epidemics. To this end, epidemiological models are commonly used. Such models study the effects of preventive measures *in silico* [5, 18].

Epidemiological models are frequently used to assist the decision making concerning the mitigation of ongoing epidemics (not only for influenza, e.g., the H1N1/09 influenza pandemic [42], but also the 2014-2016 Ebola epidemic [2], the 2016 yellow fever outbreak [24], etc.). Therefore, rapidly identifying the most promising preventive strategies is crucial. This however, can be at odds with the accuracy of the models.

There are two main types of epidemiological models that are frequently applied: *compartment models*, which divide the population into discrete homogeneous states (i.e., compartments) and describe the transition rates from one state to another, and *individual-based models* that explicitly represent all individuals and their connections, and simulate the spread of a pathogen among these individuals. While individual-based models are usually associated with a greater model complexity and computational cost than compartment models, they allow for a more accurate evaluation of preventive strategies [9, 14, 28]. It is therefore highly preferable to use individual-based models whenever computational resource constraints permit. In order to make it feasible to use individual-based models, it is essential to use the available computational resources as efficiently as possible.

The outcome of the simulation of a preventive strategy in a stochastic individual-based model, is a sample of that strategy's *outcome distribution*. In the literature, a set of possible prevention strategies is typically evaluated by simulating each of the strategies a predefined number of times (e.g., [16]). However, this can allocate a large proportion

of computational resources to explore the effects of highly sub-optimal strategies.

We therefore propose to apply *reinforcement learning* [39] with *multi-armed bandits* [3]. Reinforcement learning is the study of how to balance exploitation (i.e., further simulating the effects of what we believe to be the best preventive strategy to obtain more accurate results) and exploration (i.e., simulating the effects of other strategies to see whether they might actually be better than our current best). By using this framework, we aim to reduce the number of required model evaluations to determine the most promising preventive strategies. This reduces the total time required to study a given set of prevention strategies, making the use of individual-based models attainable in studies where it would otherwise not be computationally feasible. Additionally, faster evaluation can also free up computational resources in studies that already use individual-based models, capacitating researchers to explore different model scenarios. Considering a wider range of scenarios increases the confidence about the overall utility of prevention strategies.

In this paper, we formulate the evaluation of preventive strategies as a multi-armed bandit learning problem in section 3. The utility of this new method is confirmed through experiments in the context of pandemic influenza in section 4, using the popular FluTE individual-based model [9]. Our results show that we can quickly focus our computational resources on the optimal prevention strategy. We thus conclude that our method has the potential to be used as a decision support tool for mitigating influenza epidemics.

2. BACKGROUND

This section provides background on the application domain (i.e., finding mitigation strategies for pandemic influenza using epidemiological models) and learning methods (i.e., multi-armed bandits) approached in this study.

2.1 Pandemic influenza

Influenza is an infectious disease caused by the influenza virus. The primary prevention strategy to mitigate seasonal influenza is to produce vaccine prior to the epidemic, anticipating the virus strains that are expected to circulate. This vaccine pool is used to inoculate the population before the start of the epidemic. While seasonal influenza is usually limited to local epidemics, influenza can become pandemic when a novel virus emerges that is able to spread easily among human hosts worldwide [31]. Pandemic influenza occurs much less frequently than seasonal influenza (i.e., there were only 3 established pandemics in the 20^{th} century) but the outcome with respect to morbidity and mortality can be much more severe, potentially killing millions of people worldwide [33]. As influenza viruses are constantly evolving, the stockpiling of vaccine to prepare for a pandemic is not possible, as the vaccine should be specifically tailored to the virus that is the source of the pandemic [32]. Therefore, before an appropriate vaccine can be developed, the responsible virus needs to be identified [32]. Hence, vaccine will be available only in limited supply at the beginning of the pandemic [32]. Additionally, vaccine shortage can be induced by problems with vaccine production (e.g., the vaccine contamination in the United States in 2004-2005 [13]). While pandemic influenza has been studied and modeled extensively, there are still many aspects with respect to mitigation strategies that remain to be investigated [6, 16]. Furthermore, awareness was raised recently about certain parameters and assumptions used in epidemiological models to be too conservative to explore the full epidemiological potential of pandemic influenza, and as a result evaluate mitigation strategies overly optimistic [29]. These concerns indicate that the reevaluation of preventive strategies, taking into account more realistic assumptions, is warranted.

The severity of pandemic influenza, the limited availability of vaccine and an extensive set of open research questions renders this field a primary target to evaluate preventive strategies more efficiently.

2.2 Epidemiological models

Epidemiological models are an indispensable tool to investigate how pathogens spread through a population and to evaluate mitigation strategies. Epidemiological models are therefore crucial tools to assist policy makers with their decisions [17, 25]. Modeling epidemiological processes can be approached by means of individual-based models or compartment models. Compartment models divide the population into discrete homogeneous states (i.e., compartments) and describe the transition rates from one state to another. Compartment models can be formulated as differential equations and thus form a mathematical framework to model epidemics. Individual-based models, on the other hand, explicitly represent all individuals and their connections and simulate the spread of a pathogen among this network of individuals. Individual attributes influence the way the contact network evolves temporally and spatially. Additionally, the infection progress and the different stages associated with this progress is modeled per individual. Individual-based models allow to evaluate therapeutic and preventive interventions on the level of individuals. Compartment models generalize on population level and represent the expectation of epidemiological outcomes, while individual-based models are capable to represent individual heterogeneity. Modeling a greater level of heterogeneity is usually associated with a greater model complexity and computational cost, but allows for a more accurate evaluation of preventive strategies [9, 14, 28, 35, 41]. The result of a model evaluation is referred to as the model outcome. The relevant model outcomes greatly depend on the policy makers' research questions (e.g., prevalence, proportion of symptomatic individuals, morbidity, mortality, cost).

2.3 Modeling influenza

There is a long tradition to use individual-based models to study influenza epidemics [5, 18, 16], since it allows for a more accurate evaluation of preventive strategies. A main example is FluTE [9], an influenza individual-based model that has been the driver for many high impact research efforts over the last decade [5, 18, 21]. FluTE implements a contact model where the population is divided into communities of households [9]. The population is thus organized in a hierarchy of social mixing groups where the contact intensity is inversely proportional with the size of the group (e.g., closer contact between members of a household than between colleagues). FluTE also supports worker's commute and the travel of individuals, both model components that can be parameterized from census data. FluTE's contact network can be informed by population census data, and geographical regions as large as the United States can be modeled [9]. Next to the social mixing model, FluTE

implements an individual disease progression model, where different disease stages are associated with different levels of infectiousness. To support the evaluation of prevention strategies, FluTE allows the simulation of both therapeutic interventions (i.e., vaccines, antiviral compounds) and non-therapeutic interventions (i.e., school closure, case isolation, household quarantine). FluTE is a highly customizable simulator in which all model components can be configured in great detail.

2.4 Multi-armed bandit

The multi-armed bandit problem [22] concerns a k-armed bandit (i.e., a slot machine with k levers) where each arm A_i returns a reward r_i when it is pulled. As each arm returns rewards according to a particular reward distribution, a gambler wants to play a sequence of arms to maximize her/his reward. A strategy to play such a sequence of arms is called a *policy*. Such policies need to carefully balance between exploitation (i.e., choose the arms with the highest expected reward) and exploration (i.e., explore the other arms to potentially identify even more promising arms).
Multi-armed bandits have been proven useful to model many empirical cases: i.a., the organization of clinical trials such that patient mortality is minimized [34], resource allocation among competing stakeholders [19], adaptive routing [4], A/B testing [23] and automated auctioning [7].
The simplest policy that attempts to balance the exploitation/exploration trade-off is the ε-greedy policy [39], this policy selects the greedy arm (i.e., the arm with the highest expected reward) with probability $1-\varepsilon$ and explores the non-greedy arms with probability ε. Another popular policy is UCB1 (i.e., Upper Confidence Bound) [3]. UCB1 considers the uncertainty of each arms' value (i.e., the uncertainty of the expected reward) by selecting the arm with the highest upper confidence bound. The upper confidence bound for an arm A_i is computed as $\bar{x_i} + \sqrt{c\frac{ln(n)}{n_i}}$ where $\bar{x_i}$ is the sample average of A_i, n_i is the number of times A_i was played and n is the overall number of plays [3]. The second term is an exploratory term, which decreases when arm A_i is being pulled sufficiently. This promotes the exploration of arms for which the estimated expected reward is uncertain.

3. METHODS

To optimize the evaluation of prevention strategies, it is important to identify the best strategy using a minimal amount of model evaluations. Therefore, we propose to formulate the evaluation of prevention strategies as a multi-armed bandit problem. The presented method is generic with respect to the kind of epidemic that is modeled (i.e., pathogen, contact network, preventive strategies). The method is evaluated in the context of pandemic influenza in the next section.

3.1 Preventive bandits

Definition 1. A multi-armed bandit problem [3] consists of $n = |\{A_0, ..., A_n\}|$ arms and a (time-independent) reward distribution $P(r|A_i, \theta_i)$ for each arm, where θ_i are the parameters of the distribution. At each time step, t, an agent (i.e., gambler) chooses and plays an arm A_i, and receives a reward, r_t sampled (independently) from $P(r|A_i, \theta_i)$. The reward distributions' parameters are unknown to the agent.

The goal in a multi-armed bandit is to optimize the cumulative sum of rewards. In order to do so, it must select arms that exploit its current knowledge about θ_i, i.e., by picking the best arm it has seen so far. However, it must also explore, in order to discover arms that are better. Because the rewards are received stochastically, the agent must never exclude the possibility that its current estimates are wrong.

In our setting, we want to find the optimal preventive strategy from a set of strategies by evaluating the strategies in an epidemiological model.

Definition 2. A *stochastic epidemiological model* E is a function $\mathcal{C} \times \mathcal{P} \rightarrow \mathbb{R}$ where: $c \in \mathcal{C}$ is a configuration, $p \in \mathcal{P}$ is a preventive strategy and the codomain $\mathbb{R}$ represents the model outcome distribution.

Note that a model configuration $c \in \mathcal{C}$ describes the entire model environment. This means both aspects inherent to the model (e.g., FluTE's mixing model) and options that the modeler can provide (e.g., population statistics, vaccine properties, basic reproduction number).

Our objective is to find the optimal preventive strategy from a set of alternative preventive strategies $\{p_0, ..., p_n\} \subset \mathcal{P}$ for a particular configuration $c_0 \in \mathcal{C}$ (corresponding to the studied epidemic) of a stochastic epidemiological model. To this end, we define a preventive bandit.

Definition 3. A preventive bandit has $n = |\{p_0, ..., p_n\}|$ arms. Playing arm p_i corresponds to evaluating $E(c_0, p_i)$ by running a simulation in the epidemiological model. Evaluating $E(c_0, p_i)$ results in a sample of the model outcome distribution: *oc*. The reward of p_i is a mapping of *oc* (i.e., a sample of the model outcome distribution) using a mapping function $\mathbb{R} \rightarrow \mathbb{R}$.[1]

A preventive bandit is thus a multi-armed bandit, in which the arms are preventive strategies, and the reward distribution is implemented by an instance of a stochastic epidemiological model $E(c_0, p_i)$. We note that while the parameters of the reward distribution are in fact known, it is intractable to determine the optimal reward analytically from the stochastic epidemiological model.

Formulating the evaluation of preventive strategies in terms of a bandit problem provides us with a new framework to reason about this task. The goal is to determine the best preventive strategy (i.e., bandit arm) using as little model evaluations as possible (i.e., a best-arm identification problem).

3.2 Identifying the optimal strategy

Our goal is to identify the optimal strategy for a particular configuration $c_0 \in \mathcal{C}$ (i.e., to identify the best arm) while thoroughly exploring all preventive strategies. For this purpose, we explore the use of the popular $\varepsilon-$greedy and UCB1 algorithms.

4. EXPERIMENTS

Two experiments were composed and performed in the context of pandemic influenza modeling. More specifically, in these experiments we analyze the mitigation strategy to

[1]The mapping function allows the model outcome to be represented more conveniently for learning.

vaccinate a population when only a limited number of vaccine doses is available (details about this scenario in section 2). The experiments are inspired by the work of Medlock [27].

When the number of vaccine doses is limited, it is imperative to identify an optimal vaccine allocation strategy [27]. In our experiments, we explore the allocation of vaccines over five different age groups: pre-school children, school-age children, young adults, older adults and the elderly.

The experiments share a base model configuration, but differ with respect to a key epidemiological parameter: the basic reproduction number (i.e., R_0). The basic reproduction number represents the number of infections that is, by average, generated by one single infection.

4.1 Influenza model and configuration

The epidemiological model used in the experiments is the FluTE stochastic individual-based model (for details please refer to Appendix A). FluTE comes with a set of sample populations, in this experiment we use the sample population that describes a single community consisting of 2000 individuals (for details please refer to Appendix A). At the first day of the simulated epidemic, 10 random individuals are infected (i.e., 10 infections are seeded). The epidemic is simulated for 180 days. During this time no more infections are seeded. Thus, all new infections established during the run time of the simulation, result from the mixing between infectious and susceptible individuals. We assume no pre-existing immunity towards the circulating virus variant. We assume there are 100 vaccine doses to allocate (i.e., vaccine for 5% of the population).

In this experiment, we explore the efficacy of different vaccine allocation strategies. We consider that only one vaccine variant is available in the simulation environment. FluTE allows vaccine efficacy to be configured on 3 levels: efficacy to protect against infection when an individual is susceptible (i.e., VE_{Sus}), efficacy to avoid an infected individual from becoming infectious (i.e., VE_{Inf}) and efficacy to avoid an infected individual from becoming symptomatic (i.e., VE_{Sym}). In our experiment we consider $VE_{Sus} = 0.5$ [26], $VE_{Inf} = 0.5$ [26] and $VE_{Sym} = 0.67$ [42]. The influenza vaccine, as most vaccines, only becomes fully effective after a certain period upon its administration, and the effectiveness increases gradually over this period [1]. In our experiment, we assume the vaccine effectiveness to build up linearly over a period of 2 weeks [1, 8].

We define two experiments: both experiments use the base model configuration as described above. The two experiments differ with respect to their R_0 (i.e., basic reproduction number) parameter. To evaluate our new method, we select 2 values that are used in many studies: $R_0 = \{1.3, 1.4\}$ [5, 9, 27]. Each experiment thus has its own configuration. With respect to the definition of the epidemiological model (i.e., $E = \mathcal{C} \times \mathcal{P} \rightarrow \mathbb{R}$), we can express these configurations as $c_{R_0=1.3}$ and $c_{R_0=1.4} \in \mathcal{C}$.

4.2 Formulating vaccine allocation strategies

We consider 5 age groups to which vaccine doses can be allocated: pre-school children (i.e., 0-4 years old), school-age children (i.e., 5-18 years old), young adults (i.e., 19-29 years old), older adults (i.e., 30-64 years old) and the elderly ($>$ 65 years old). An allocation scheme can be encoded by a Boolean 5-tuple, where each position in the tuple corresponds to the respective age group. When the value is 1 at a position, this denotes that vaccines should be allocated to the respective age group. When the value is 0 at a position, this denotes that vaccines should not be allocated to the respective age group. When vaccine is to be allocated to a particular age group, this is done proportional to the size of the population that is part of this age group.

Some examples: a preventive strategy where no vaccine should be allocated is encoded as $\langle 0, 0, 0, 0, 0 \rangle$, a preventive strategy where vaccine needs to be allocated uniformly across all age groups is encoded as $\langle 1, 1, 1, 1, 1 \rangle$, a preventive strategy where vaccine needs to be allocated exclusively to children is encoded as $\langle 1, 1, 0, 0, 0 \rangle$.

To decide on the best vaccine allocation strategy, we enumerate all possible combinations of this tuple. Since the tuple consists of a sequence of $\{0, 1\}^*$, the tuple can be encoded as a binary number. This enables us to represent the different allocation strategies by integers (i.e., $\{0, 1, ..., 31\}$).

With respect to the definition of the epidemiological model (i.e., $E = \mathcal{C} \times \mathcal{P} \rightarrow \mathbb{R}$), this set of 32 strategies is a subset of $\mathcal{P}$.

4.3 An influenza bandit

So far, we defined the model configurations (i.e., $c_{R_0=1.3}$ and $c_{R_0=1.4}$) and the set of preventive strategies (i.e., 32 vaccine allocation strategies) to be evaluated.

Now, let us define the *influenza preventive bandit* B_{Flu}: B_{Flu} has exactly 32 arms (i.e., $\{A_0, ..., A_{31}\}$). Each arm A_i is associated with the allocation strategy for which the integer encoding is equal to i. To conclude the specification of the influenza bandit B_{Flu}, we describe what happens when an arm A_i of B_{Flu} is played:

1. Invoke FluTE with a model configuration $c_0 \in \mathcal{C}$ and the vaccine allocation strategy $p_i \in \mathcal{P}$ associated with the arm A_i (i.e., this is allocation strategy i, using the strategy's integer representation).[2]

2. From FluTE's output, extract the proportion of the population that experienced a symptomatic infection: $\frac{\#\text{ symptomatic individuals}}{\#\text{ individuals}}$.

3. Return a $reward = 1 - \frac{\#\text{ symptomatic individuals}}{\#\text{ individuals}}$. Note that the reward denotes the proportion of individuals that did not experience symptomatic infection.

4.4 Outcome distributions

To perform an initial analysis concerning the outcome distributions of the 32 prevention strategies, all strategies were evaluated 1000 times for both model configurations (i.e., $c_{R_0=1.3}$ and $c_{R_0=1.4} \in \mathcal{C}$). Note that generating thousands of samples (i.e., 2×32000 in this case) would not be computationally feasible when considering a larger population. This analysis is performed to identify the best strategy, such that we can properly validate the results from our learning experiments.

The outcome distributions are visualized in Figure 1 and Figure 2 for $c_{R_0=1.3}$ and $c_{R_0=1.4}$ respectively. A violin plot is used to plot the density of the outcome distribution per vaccine allocation strategy. The density for a particular strategy is computed based on 1000 samples of the strategy's outcome distribution. Note that while the distributions have

[2]Note that the configuration is serialized as a text file, for details on the format of this file, refer to Appendix B.

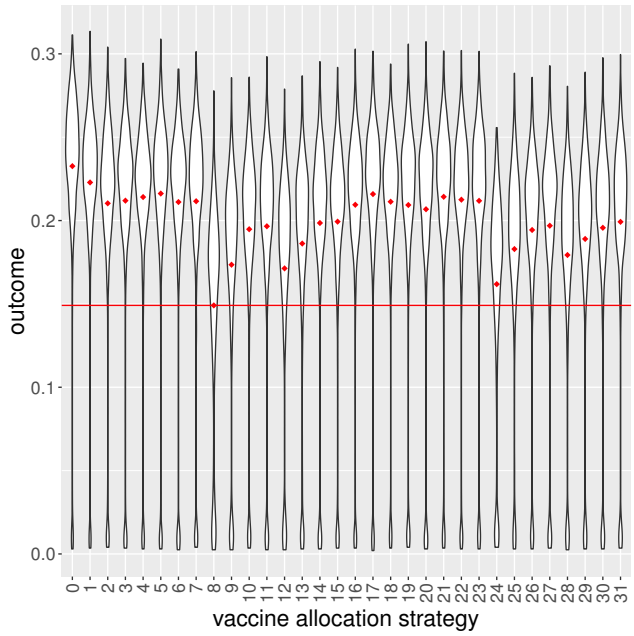

Figure 1: Violin plot that depicts the density of the outcome distribution for 32 vaccine allocation strategies, considering a model environment with $R_0 = 1.3$. For each density, the sample mean is visualized with a red diamond. The sample mean of the optimal strategy is depicted with a red horizontal line.

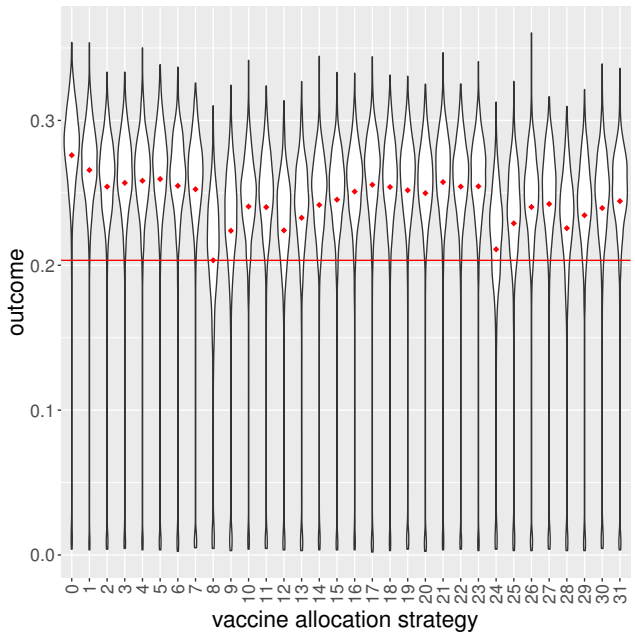

Figure 2: Violin plot that depicts the density of the outcome distribution for 32 vaccine allocation strategies, considering a model environment with $R_0 = 1.4$. For each density, the sample mean is visualized with a red diamond. The sample mean of the optimal strategy is depicted with a red horizontal line.

considerable density around the mean of the distribution, there is also quite some density where the outcome is close to 0. This is an artefact of the stochastic simulation: the pathogen is not able to establish an epidemic for certain simulation runs.

Our analysis shows that the best vaccine allocation strategy was identified to be $\langle 0, 1, 0, 0, 0 \rangle$ (i.e. vaccine allocation strategy 8) for both model configurations $c_{R_0=1.3}$ and $c_{R_o=1.4}$.

4.5 UCB1 and ε-greedy experiment

To explore the utility of bandits to evaluate preventive strategies, we average over 500 independent bandit runs for both experiments. For each experiment, we run the ε-greedy ($\varepsilon = 0.1$) and UCB1 algorithm for 1000 iterations [3].

The average reward reported in the first experiment is visualized in Figure 3 for both the $\varepsilon-$greedy and UCB1 algorithm. The average reward reported in the second experiment is visualized in Figure 4 for both the $\varepsilon-$greedy and UCB1 algorithm.

We observe that the average reward starts to increase from iteration 400, for both $\varepsilon-$greedy and UCB1, and continues to increase for the rest of the iterations. However, we also note that the average reward learning curve increases faster for ε-greedy than for UCB1.

In the previous section, the best vaccine allocation strat-

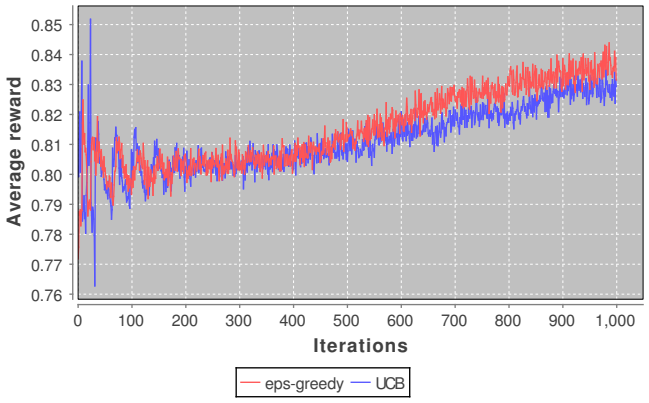

Figure 3: Reward learning curve for the first experiment (i.e., model with $R_0 = 1.3$), averaged over 500 independent bandits for 1000 iterations. This plot depicts the learning curve for both the $\varepsilon-$greedy (i.e., red curve) and UCB1 (i.e., blue curve) algorithms.

[3]To remind the reader, each arm involves the invocation of the FluTE simulator, and is therefore associated with a significant computational cost (for details, please see Appendix D).

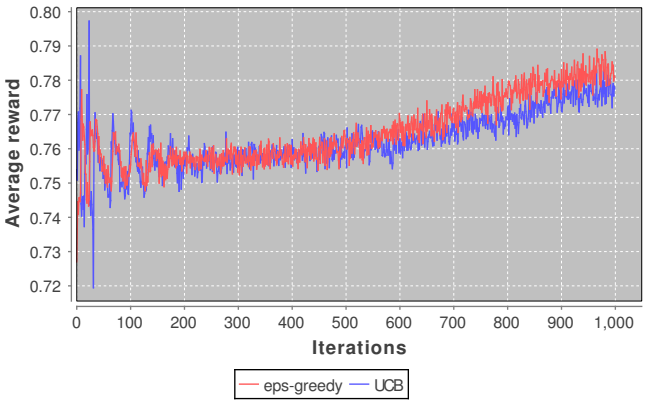

Figure 4: Reward learning curve for the first experiment (i.e., model with $R_0 = 1.4$), averaged over 500 independent bandits for 1000 iterations. This plot depicts the learning curve for both the $\varepsilon-$greedy (i.e., red curve) and UCB1 (i.e., blue curve) algorithms.

egy was identified to be $\langle 0, 1, 0, 0, 0\rangle$ (i.e. vaccine allocation strategy 8) for both $c_{R_0=1.3}$ and $c_{R_o=1.4}$. Figure 5 visualizes the percentage of plays of the optimal arm (i.e., vaccine allocation strategy $\langle 0, 1, 0, 0, 0\rangle$) for the first experiment. Figure 6 visualizes the percentage of plays of the optimal arm (i.e., vaccine allocation strategy $\langle 0, 1, 0, 0, 0\rangle$) for the second experiment.

For both of the experiments, $\varepsilon-$greedy ends up selecting optimal actions 60% of the time after 1000 iterations. As we observed for the average reward learning curve, UCB1 also performs worse with respect to the optimal action selection learning curve, reaching only 40-45% optimal action selection.

5. DISCUSSION

Our influenza model, and more specifically the context in which only a limited set of vaccine doses is available, was inspired by the work presented by Medlock [27]. However, we consider a much smaller population (i.e., 2000 individuals versus the entire United States), to make it computationally feasible to validate our learning experiments. Furthermore, because of the differences between the model setup presented by Medlock and FluTE, a perfect mapping was not possible. It would therefore not be sound to compare our results directly to the results obtained by Medlock. We were, however, able to reproduce some significant trends. The best strategy identified by our analyses is associated with the allocation of vaccine to children: this is in agreement with Medlock's work.

The analysis of the outcome distributions for the different vaccine allocation strategies shows that there is one optimal strategy $\langle 0, 1, 0, 0, 0\rangle$. The differences between the means and medians of the different strategies are however not very pronounced. This is related to the limited number of available vaccine doses. While this is a setting that would not be majorly promising to policy makers, it provides us with an interesting environment to test the performance of the preventive bandits.

For both of the experiments, $\varepsilon-$greedy ends up selecting optimal actions 60% of the time after 1000 iterations. These

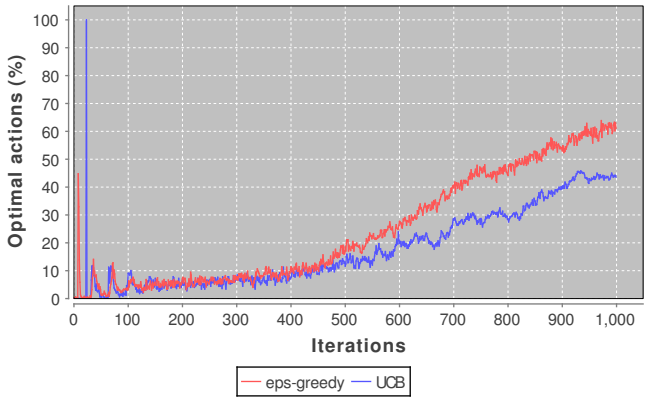

Figure 5: Optimal action selection learning curve for the first experiment (i.e., model with $R_0 = 1.3$), averaged over 500 independent bandits for 1000 iterations (i.e., the Y-axis depicts the % the optimal action was selected). This plot depicts the learning curve for both the $\varepsilon-$greedy (i.e., red curve) and UCB1 (i.e., blue curve) algorithms.

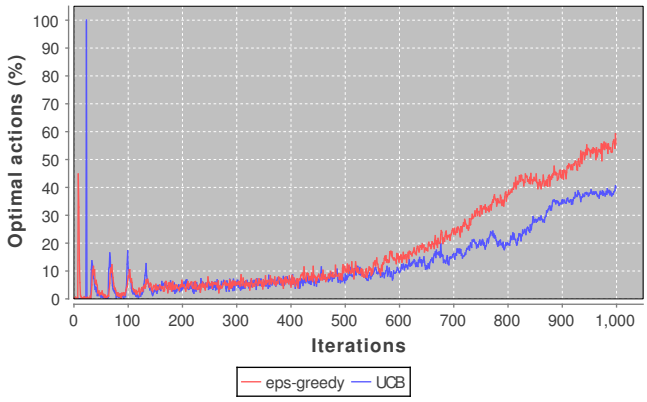

Figure 6: Optimal action selection learning curve for the first experiment (i.e., model with $R_0 = 1.4$), averaged over 500 independent bandits for 1000 iterations (i.e., the Y-axis depicts the % the optimal action was selected). This plot depicts the learning curve for both the $\varepsilon-$greedy (i.e., red curve) and UCB1 (i.e., blue curve) algorithms.

results demonstrate that it is possible to identify the optimal strategy using only a limited number of model evaluations, even if there is a large number of preventive strategies to consider. We also observe, that both the average reward and optimal action selection learning curves continue to increase, indicating that the learning has not yet converged. It is however important to stress that, our main interest is not convergence, but to identify the best strategy using a minimal number of model evaluations.

We observe that, in our experiment setting, $\varepsilon-$greedy outperforms UCB1, both with respect to the average reward learning curve and the optimal action selection learning curve.

To support the reproducibility of our research, all source code and configuration files used in our experiments is publicly available (for details, please see the Appendices).

6. CONCLUSIONS

We formally defined the evaluation of prevention strategies as a multi-armed bandit problem. We used this formal definition to describe a bandit that can be used to evaluate vaccine allocation strategies with the intention to mitigate pandemic influenza. Two elaborate experiments were set up to evaluate this preventive bandit using the popular FluTE individual-based model. To assess the performance of the preventive bandit, we report an average over 500 independent bandit runs, for the two experiments.

We demonstrate that it is possible to identify the optimal strategy using only a limited number of model evaluations, even if there is a large number of preventive strategies to consider.

We are confident that our method has the potential to be used as a decision support tool for mitigating influenza epidemics. To increase this potential, we aim to significantly extend the features of our tool and framework.

Firstly, while our method is evaluated in the context of pandemic influenza, it is important to stress that both our formalisms and infrastructure can be used to evaluate prevention strategies for other infectious diseases. We expect that epidemics of arboviruses (i.e., viruses that are transmitted by a mosquito vector; e.g., Zika virus, Dengue virus) are a particularly interesting use case for our preventive bandits. Only since recently, Dengue and Zika vaccines are available [20] or in the pipeline [11], and the optimal allocation of these vaccines is an important research topic [15]. Additionally, there exist individual-based arbovirus models [10] that could be readily applied to perform such analyses. We aim to test our approach on these pathogens as well.

Secondly, we aim to make different algorithmic extensions. In this study, we used elemental bandit learning algorithms (i.e., $\varepsilon-$greedy and UCB1). We acknowledge that other algorithms could potentially learn faster. We created the infrastructure to easily implement and experiment with different algorithms and epidemiological models (details can be found in the Appendices) and we will use this framework to explore the use of other algorithms. Furthermore, the use of stateless reinforcement learning (i.e., bandits) presents us with a stepping stone to consider reinforcement learning where the partial or full state of the epidemiological model (e.g., which people are currently infected, and which measures have already been taken and to what effect) is used to learn preventive strategies that are more reactive towards events that take place in the simulation. We believe that such strategies may prove to be better than the static strategies we used in this study.

Finally, our current preventive bandits only learn with respect to a single model outcome: more specifically, for influenza this is the proportion of symptomatic infections. In the context of influenza, and for many infectious diseases, there is often interest to consider additional model outcomes (e.g., morbidity, mortality, cost). In the future, we aim to use *multi-objective multi-armed bandits* [12] in contrast to the current single-objective preventive bandits. With this approach, we plan to learn a *coverage set* containing an optimal strategy for every possible preference profile the decision makers might have [37]. We aim to design suitable quality metrics [36, 40, 43] tailored to the use case of epidemiological preventive strategy learning, to support the entire spectrum of epidemiological models and thus to prevent method overfitting [43].

Acknowledgments

Pieter Libin was supported by a PhD grant of the FWO (Fonds Wetenschappelijk Onderzoek − Vlaanderen) and the VUB research council (VUB/OZR2714). Timothy Verstraeten was supported by a PhD grant of the FWO (Fonds Wetenschappelijk Onderzoek − Vlaanderen) and the VUB research council (VUB/OZR2884). Kristof Theys was supported by a postdoctoral grant of the FWO (Fonds Wetenschappelijk Onderzoek − Vlaanderen). Diederik Roijers was supported by a postdoctoral grant of the FWO (Fonds Wetenschappelijk Onderzoek − Vlaanderen). Thanks to Roxana Rădulescu, for her careful proofreading and helpful suggestions. Thanks to Karen Goedeweeck, for her careful proofreading. Thanks to all gnu/python/R/Scala/FluTE hackers for their efforts, useful libraries and excellent tools. The computational resources and services used in this work were provided by the Hercules Foundation and the Flemish Government department EWI-FWO Krediet aan Navorsers (Theys, KAN2012 1.5.249.12.).

APPENDIX

A. FLUTE SOURCE

FluTE is a stochastic individual-based model, that is implemented in C++. The original source code, as release by FluTE's author (i.e., D. Chao), is available from https://github.com/dlchao/FluTE. This github repository contains FluTE's C++ source code, GNU/Linux-specific make files and a set of population density descriptions that can be used to simulate particular geographical settings (i.e., 2000-individual population, Seattle, Los Angelos and the entire United States).

Some changes were made to the source code to make our research easier: we organized the source code in a directory structure and added a CMake meta-make file. This CMake build file allows us to build the source code on GNU/Linux and MacOS [4]. These changes are publicly available on the https://github.com/vub-ai-lab/FluTE-bandits github repository.

B. FLUTE CONFIGURATIONS

[4]Microsoft Windows should also work with little changes, but this was not tested yet.

To run our experiments, we defined a model environment to evaluate pre-vaccination with little vaccine available, as described in detail in section 4. The pre-vaccination configuration script can be found in the 'configs/bandits' directory of the https://github.com/vub-ai-lab/FluTE-bandits github repository. Note that this configuration script is a python Mako template (http://makotemplates.org/), to enable easy parameterization of the configuration script.

C. BANDIT IMPLEMENTATION

We implemented a flexible bandit framework in Scala, the code is publicly available on github: https://github.com/vub-ai-lab/scala-bandits. This framework is specifically designed to enable us to easily experiment with new algorithms and environments (i.e., both Scala environments and external environments, such as e.g., the FluTE simulator environment). The repository contains the $\varepsilon-$greedy algorithm, the UCB1 algorithm, the Sutton test environment [39], the FluTE environment and some post processing utilities.

D. HIGH PERFORMANCE COMPUTING

Simulating epidemics using individual-based models is a computationally intensive process. Therefore, our experiments were run on a powerful high performance computing cluster: the Flemish Supercomputer Center. We report that, to make this possible, all software had to be installed (or build) for the high performance computing cluster. We report that our FluTE CMake file allows the generation of efficient code (i.e., using SSE instructions) for all plaforms used in our analyses (i.e., MacOS, XUbuntu desktop GNU/Linux and GNU/Linux on the high performance computing cluster).

REFERENCES

[1] A. K. Abbas, A. H. H. Lichtman, and S. Pillai. *Cellular and molecular immunology*. Elsevier Health Sciences, 2014.

[2] M. Ajelli, S. Merler, L. Fumanelli, A. Pastore y Piontti, N. E. Dean, I. M. Longini, M. E. Halloran, and A. Vespignani. Spatiotemporal dynamics of the Ebola epidemic in Guinea and implications for vaccination and disease elimination: a computational modeling analysis. *BMC Medicine*, 14(1):1–10, 2016.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[4] B. Awerbuch and R. Kleinberg. Near-optimal adaptive routing: Shortest paths and geometric generalizations. In *Proceeding of the 36th Annual ACM Symposium on Theory of Computing*, pages 45–53, 2004.

[5] N. E. Basta, D. L. Chao, M. E. Halloran, L. Matrajt, and I. M. Longini. Strategies for pandemic and seasonal influenza vaccination of schoolchildren in the United States. *American journal of epidemiology*, 170(6):679–686, 2009.

[6] M. Biggerstaff, C. Reed, D. L. Swerdlow, M. Gambhir, S. Graitcer, L. Finelli, R. H. Borse, S. A. Rasmussen, M. I. Meltzer, and C. B. Bridges. Estimating the potential effects of a vaccine program against an emerging influenza pandemic - United States. *Clinical Infectious Diseases*, 60:S20–S29, 2015.

[7] A. Blum, V. Kumar, A. Rudra, and F. Wu. Online learning in online auctions. *Theoretical Computer Science*, 324(2-3):137–146, 2004.

[8] CDC. Key facts about influenza (flu) & flu vaccine. *Atlanta, GA: Centers for Disease Control and Prevention*, 2014.

[9] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, 6(1):e1000656, 2010.

[10] D. L. Chao, S. B. Halstead, M. E. Halloran, and I. M. Longini. Controlling Dengue with Vaccines in Thailand. *PLoS Neglected Tropical Diseases*, 6(10), 2012.

[11] J. Cohen. The race for a Zika vaccine is on. *Science*, 351(6273):543–544, 2016.

[12] M. M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *Proceedings of the International Joint Conference on Neural Networks*, 2013.

[13] M. Enserink. Crisis underscores fragility of vaccine production system. *Science*, 306(5695):385, 2004.

[14] S. Eubank, V. Kumar, M. Marathe, A. Srinivasan, and N. Wang. Structure of social contact networks and their impact on epidemics. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 70(0208005):181, 2006.

[15] N. M. Ferguson, I. Rodríguez-Barraquer, I. Dorigatti, L. Mier-y Teran-Romero, D. J. Laydon, and D. A. T. Cummings. Benefits and risks of the Sanofi-Pasteur dengue vaccine: Modeling optimal deployment. *Science*, 353(6303):1033–1036, 2016.

[16] L. Fumanelli, M. Ajelli, S. Merler, N. M. Ferguson, and S. Cauchemez. Model-Based Comprehensive Analysis of School Closure Policies for Mitigating Influenza Epidemics and Pandemics. *PLoS Computational Biology*, 12(1), 2016.

[17] G. P. Garnett, S. Cousens, T. B. Hallett, R. Steketee, and N. Walker. Mathematical models in the evaluation of health programmes, 2011.

[18] T. C. Germann, K. Kadau, I. M. Longini, and C. A. Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, 2006.

[19] J. Gittins, K. Glazebrook, and R. Weber. *Multi-Armed Bandit Allocation Indices: 2nd Edition*. 2011.

[20] S. R. Hadinegoro, J. L. Arredondo-García, M. R. Capeding, C. Deseda, T. Chotpitayasunondh, R. Dietze, H. H. M. Ismail, H. Reynales, K. Limkittikul, D. M. Rivera-Medina, H. N. Tran, A. Bouckenooghe, D. Chansinghakul, M. Cortés, K. Fanouillere, R. Forrat, C. Frago, S. Gailhardou, N. Jackson, F. Noriega, E. Plennevaux, T. A. Wartel, B. Zambrano, and M. Saville. Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. *New England Journal of Medicine*, 373(13):1195–1206, 2015.

[21] M. E. Halloran, I. M. Longini, A. Nizam, and Y. Yang. Containing bioterrorist smallpox. *Science (New York, N.Y.)*, 298(5597):1428–1432, 2002.

[22] R. Herbert. Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematical*

Society, 58(5):527–535, 1952.

[23] E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of A/B Testing. In *COLT*, pages 461–481, 2014.

[24] M. U. G. Kraemer, N. R. Faria, R. C. Reiner, N. Golding, B. Nikolay, S. Stasse, M. A. Johansson, H. Salje, O. Faye, G. R. W. Wint, and Others. Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. *The Lancet Infectious Diseases*, 2016.

[25] J. Lessler, W. J. Edmunds, M. E. Halloran, T. D. Hollingsworth, and A. L. Lloyd. Seven challenges for model-driven data collection in experimental and observational studies. *Epidemics*, 10:78–82, 2014.

[26] H. Q. McLean, M. G. Thompson, M. E. Sundaram, B. A. Kieke, M. Gaglani, K. Murthy, P. A. Piedra, R. K. Zimmerman, M. P. Nowalk, J. M. Raviotta, M. L. Jackson, L. Jackson, S. E. Ohmit, J. G. Petrie, A. S. Monto, J. K. Meece, S. N. Thaker, J. R. Clippard, S. M. Spencer, A. M. Fry, and E. A. Belongia. Influenza vaccine effectiveness in the United States during 2012-2013: Variable protection by age and virus type. *Journal of Infectious Diseases*, 211(10):1529–1540, 2015.

[27] J. Medlock and A. P. Galvani. Optimizing influenza vaccine distribution. *Science*, 325(5948):1705–1708, 2009.

[28] L. A. Meyers, M. E. J. Newman, M. Martin, and S. Schrag. Applying network theory to epidemics: Control measures for Mycoplasma pneumoniae outbreaks. *Emerging Infectious Diseases*, 9(2):204–210, 2003.

[29] M. A. Miller, C. Viboud, M. Balinska, and L. Simonsen. The Signature Features of Influenza Pandemics: Implications for Policy. *New England Journal of Medicine*, 360(25):2595–2598, 2009.

[30] N. A. M. Molinari, I. R. Ortega-Sanchez, M. L. Messonnier, W. W. Thompson, P. M. Wortley, E. Weintraub, and C. B. Bridges. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25(27):5086–5096, 2007.

[31] H. Nicholls. Pandemic influenza: the inside story. *PLoS Biol*, 4(2):e50, 2006.

[32] W. H. Organization and Others. WHO guidelines on the use of vaccines and antivirals during influenza pandemics. 2004.

[33] K. D. Patterson and G. F. Pyle. The geography and mortality of the 1918 influenza pandemic. *Bulletin of the History of Medicine*, 65(1):4, 1991.

[34] W. H. Press. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences*, 106(52):22387–22392, 2009.

[35] H. Rahmandad and J. Sterman. Heterogeneity and Network Structure in the Dynamics of Diffusion: Comparing Agent-Based and Differential Equation Models. *Management Science*, 54(5):998–1014, 2008.

[36] D. M. Roijers. *Multi-objective decision-theoretic planning*. PhD thesis, University of Amsterdam, 2016.

[37] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[38] K. Stöhr. Influenza: WHO cares. *The Lancet infectious diseases*, 2(9):517, 2002.

[39] R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction*. 1998.

[40] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine learning*, 84(1-2):51–80, 2011.

[41] L. Willem. *Agent-Based Models For Infectious Disease Transmission: Exploration, Estimation & Computational Efficiency*. PhD thesis, 2015.

[42] W. Yang, J. D. Sugimoto, M. E. Halloran, N. E. Basta, D. L. Chao, L. Matrajt, G. Potter, E. Kenah, and I. M. Longini. The transmissibility and control of pandemic influenza A (H1N1) virus. *Science (New York, N.Y.)*, 326(2009):729–33, 2009.

[43] L. M. Zintgraf, T. V. Kanters, D. M. Roijers, F. A. Oliehoek, and P. Beau. Quality assessment of MORL algorithms: A utility-based approach. In *Benelearn 2015: Proceedings of the Twenty-Fourth Belgian-Dutch Conference on Machine Learning*, 2015.