New features in EpiGraph for optimizing COVID-19 simulation accuracy

David E. Singh [1,*], Maria-Cristina Marinescu [2], Miguel Guzmán-Merino [1], Christian Durán [1], Concepción Delgado-Sanz [3,4], Diana Gomez-Barroso [3,4] and Jesus Carretero [1]

[1] Department Computer Science, Universidad Carlos III de Madrid, Leganés, Spain.

[2] Barcelona Supercomputing Center, Barcelona, Spain.

[3] CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

[4] National Centre for Epidemiology, Carlos III Institute of Health, Madrid, Spain

1 Introduction

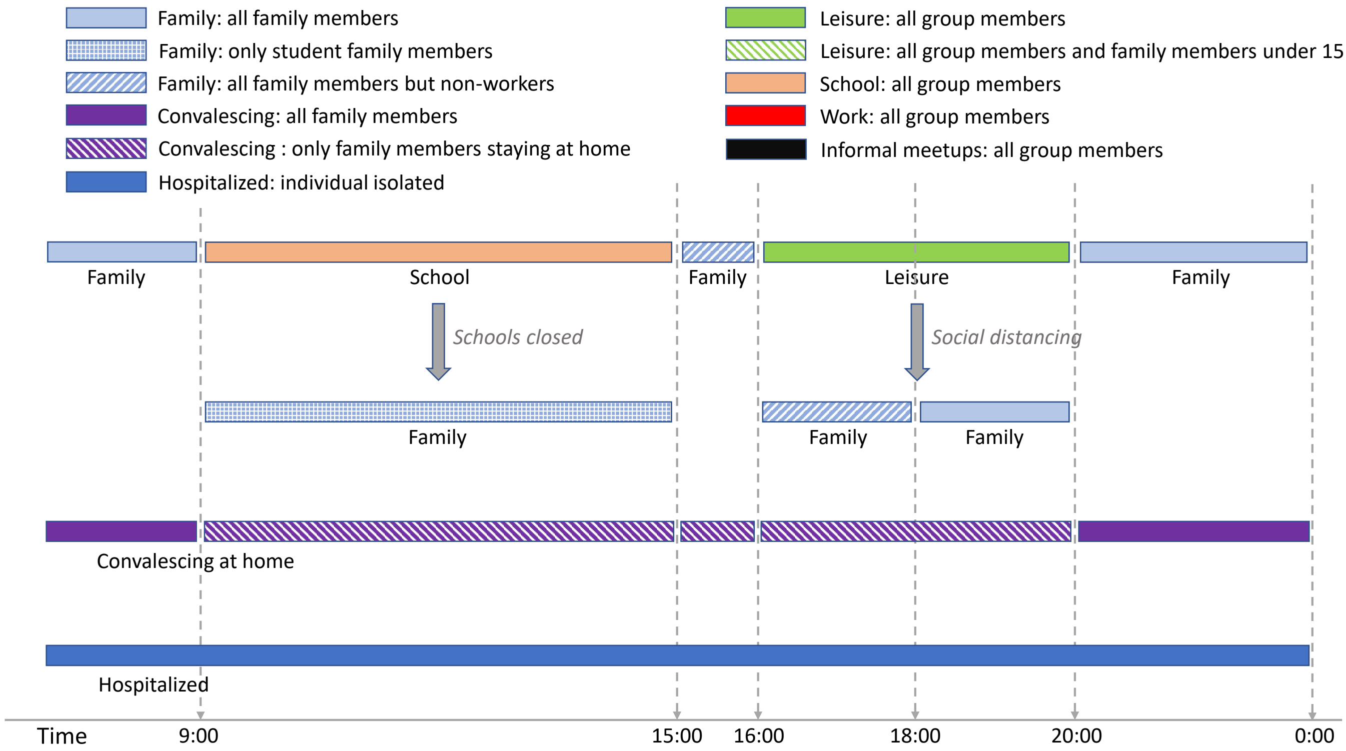
EpiGraph is a scalable, fully distributed simulator that performs large scale and realistic stochastic simulations of the propagation of the influenza and COVID19 viruses. EpiGraph, which we described in [3, 2, 6], includes an agent-based model that captures individual characteristics and specifies the interaction patterns based on existing virtual connections extracted from social networks and contact matrices. These connections are time-dependent in order to realistically capture the temporal nature of interactions and some of them are specific for each profession. This work describes the simulator implementation that includes the following new features designed to increase the simulation accuracy: (1) a more detailed social model that leverages contact matrices to determine the individual's number of contacts, which now are age-dependent, (2) the distinction between different professions, each one of them with different characteristics and iteration patterns, (3) new epidemic compartment states, (4) a new vaccination model.

2 Materials and methods

2.1 Transmission algorithm

Algorithm 1 shows an outline of EpiGraph's simulation algorithm. The iterative algorithm discretizes the total simulation time in time steps of 10 minutes (line 1). In each time step, the algorithm considers each city in the simulated territory (line 2). A city has a given population which is modelled based on the Spanish census data [4], with the associated social connections between the individuals. Line 5 updates the health status of each infected individual of each city, as indicated by the epidemic model used by the simulator. The next step (line 6) computes how the infectious agent spreads via the social model using the graphs extracted from social networks, starting from every infected individual and evaluating the probability of transmission to each one of their contacts. This probability depends, among others, on the type of connection, the time of day, and the characteristics of the individual potentially being infected, such as their age, profession or whether the infected or the susceptible individuals are using masks. In line 7 dynamic transmission is evaluated. As opposed to the graph-based transmission, these connections are dynamically generated for individuals belonging to certain professions that are in constant contact with different people (for instance, catering workers).

We call an individual intervention (line 10) an action taken by the individual to mitigate the propagation of the infectious disease. In Epigraph these actions are activated or deactivated based on defined policies. On example of intervention is that at simulation day 30, a certain individual starts using surgical face masks at work, but not at family time. We call a social intervention (line 11) those interventions —such as school closing or social distancing— that are imposed (or lifted) by the health authorities at a certain time of the simulation.



(a) Student activity cycle for weekdays

Figure 1: Student activity cycle for weekdays (a). The interaction patterns are specific for Spain. Night time is considered to extend from midnight to 9:00 and is not included in the figure (but is considered in the simulation). The text in italic shows the effect of social restrictions.

In the propagation of the infection via the Transportation model (line 12), some individuals move between their city and another, depending on the city sizes and the geographical distance between them. This allows us to model the medium and long distance travel of people. Finally, the Vaccination model includes both the COVID-19 vaccine characteristics and the time in which the individuals are vaccinated -including the vaccine type that they receive-.

Algorithm 1 EpiGraph transmission algorithm. Variable *simulation_time* represents the simulation duration, *simulated_territory* is the simulated area including several cities, each one of them with a social interaction model for the population, and *status* contains characteristics and health status of each individual for each city.

1: **for** $timestep = 1 \rightarrow simulation\_time$ **do**
2:   **for** $city \in simulated\_territory$ **do**
3:     **for** $individual \in city$ **do**
4:       **if** $status[individual]$ is $infectious$ **then**
5:         $UpdateStatus(status[individual])$
6:         $ComputeSpreadGraph(individual, city)$
7:         $ComputeSpreadDynamic(individual, city)$
8:       **end if**
9:     **end for**
10:     $Individual\_Interventions(status)$
11:     $Social\_Interventions(city)$
12:     $Transportation(city, simulated\_territory)$
13:     $Vaccination(city)$
14:   **end for**
15: **end for**

2.2 Social model

EpiGraph's social model is an agent-based model that captures individual attributes and specifies the way that the individuals interact based on patterns extracted from social networks (Facebook) and from companies (Enron Email Corpus) and is also combined with the information of contact matrices that define the average number of contacts per age interval for each country. We use demographic information to reproduce social habits for four different main collectives: students, workers, stay-at-home people, and elders. For each collective, the individuals are gathered in groups. One group represents a certain number of individuals that interact during the work time. For instance, groups are the students belonging to the same classroom, workers of the same company and stay-at-home people or elderly people that perform group activities. Table 1 shows the size of each collective including the amount of groups related to each one. Work and elder collectives are broken down in different professions and elderly types. Tables 2 and 3 shows the different professions and groups considered for these collectives. The percentages of each collective correspond to the average values of Spanish society [4].

The way the individuals establish social contacts[1] is time-dependent in order to realistically reflect the temporal nature of the different classes of interactions that each individual has throughout the day. For each one of the group types we consider three different temporal distribution of the individual's activities, those related to weekdays, Saturdays, and holidays (including Sundays).

Figure 1 shows the activity cycles for students during weekdays. Similar activity cycles are included for weekends and holidays and for the other collectives. A more detailed description of these activities can is depicted in [6]. These patterns are specific to the place being modelled; in Spain, for instance, breakfast is around 8:00, lunch time around 14:00, and dinner time starts at 20:00. The period ranging from 0:00 until 9:00 (not shown in the figures)

[1] We define a social contact (also called contact or interaction) between two individuals a co-location in time (and space) at a distance that is small enough to make disease transmission possible.

Collective	# Groups	# Individuals
Workers	13,763	108,188
Students	289	35,573
Elderly	3,972	38,889
Unemployed	2,925	17,350

Table 1: Number of groups and individuals for each collective for a simulated city with of 500,000 inhabitants .

	Industry	**Building**	**Catering**	**Services**	**Security**	**Edu.**	**Health**	**Elderly-CG**	**Transport**
	30.80%	6.50%	8.80%	24.00%	7.40%	7.50%	6.40%	3.30%	5.30%
$Size_{min}$	1	1	1	1	10	6	10	5	1
$Size_{max}$	30	20	12	8	50	30	30	25	8

Table 2: Work collective breakdown in professions. Edu. and Elderly-CG stands for Education and Elderly caregiver, respectively. The percentages are the fraction of each profession among the worker collective. $Size_{min}$ and $Size_{max}$ denote the minimum and maximum sizes of each specific collective. A random uniform distribution between these two values has been used for setting each group size.

	Elderly at home	**Elderly at day-care centre**	**Elderly at nursing home**
	50.6%	46.3%	3.1%
$Size_{min}$	4	10	10
$Size_{max}$	10	30	40

Table 3: Elderly collective breakdown in classes. Elderly at home represents the elderlies that live at home and participate in day centres (note that in our simulations day centres are closed when the social distancing policies are applied). The percentages are the fraction of each profession among the elderlies. $Size_{min}$ and $Size_{max}$ denote the minimum and maximum sizes of each specific collective. A random uniform distribution between these two values has been used for setting each group size.

corresponds to family time. Note that family time includes all the activities carried out at home (dinner time, family time and night sleep). We assume that school time is followed by a short period of family time, after which there is a leisure period in which the students are in contact with other individuals different from those belonging to the same school group. There are two social distancing policies applied for students: school closure, and social distancing, in which school and leisure times are replaced by family time. In EpiGraph we distinguish between different levels of family-time interactions, based on the family members that are at home at each time of the day as follows: at night-time each individual is in contact with all family members; when schools are closed, the family-time for this period takes into account only those family members that are at home. For instance, if work places are opened, family time will not include the working members. On the other hand, when social distancing is not imposed, social contacts with stay-at-home and elderly family members are not taken into account during this time period because we assume that these two group types are not at home at this time. For the same reasons, during the family time slot from 15:00 to 16:00, stay-at-home persons and elderly family members are included in the interactions, while working members are not.

EpiGraph creates different graphs for each work group, school group, stay-at-home (informal meetup) group, and elderly (informal meetups) group. Rather than assuming a distribution or generating synthetic interaction graphs, we use real information from social networks to model the social interaction patterns. Each group has a different size, in between $Size_{min}$ and $Size_{max}$. We have used the Enron Email Corpus (70,578 nodes and 312,620 edges) for generating the work, elderly and informal-meetup groups while the Facebook (250,000 edges and 3,239,137 edges) network was used to generate the school groups. We designed a graph-scaling algorithm that selects as many nodes as

Ages	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90+**
90+	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0
80-90	0.3	0.3	0.4	0.4	0.7	0.6	0.6	0.9	0.5	0.1
70-80	0.3	0.3	0.6	1.1	1	0.8	1	1	0.5	0.1
60-70	0.3	0.4	1.5	1.8	1.4	1.2	1	0.9	0.3	0.1
50-60	0.5	1.1	2.3	2.6	2	1.5	0.9	0.5	0.2	0.1
40-50	1	1.4	2.5	2.8	2.3	1.5	0.8	0.5	0.2	0.1
30-40	1.3	1.3	2.5	3.2	2.4	1.6	0.9	0.5	0.1	0.1
20-30	1	1.4	4	3	2.6	1.8	0.9	0.3	0.1	0.1
10-20	2.1	8.3	1.9	2.1	1.8	1.1	0.3	0.2	0.1	0.1
0-10	7.9	2.1	1.2	1.9	1.3	0.4	0.3	0.2	0.1	0.1
Total	**14.7**	**16.7**	**17**	**19.1**	**15.7**	**10.8**	**6.9**	**5.1**	**2.2**	**0.5**

Table 4: Global input contact matrix used in the social model. Each column represents the average number of contacts of the individuals belonging to the age interval shown on the column's top. These contacts are broken down by age intervals in the different rows.

Ages	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90+**
90+	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.4	0.4	0.3
80-90	0.2	0.2	0.3	0.4	0.3	0.3	0.4	1	0.6	0.6
70-80	0.4	0.4	1.3	1.4	1.2	1.1	1	1.3	1	1
60-70	0.5	1.2	2.2	2.1	1.6	1.5	0.8	0.7	0.2	0.2
50-60	1.2	1.7	2.7	2.7	2.3	1.8	1	0.8	0.2	0.2
40-50	1.6	1.7	2.8	3.1	2.1	1.8	0.9	0.8	0.2	0.2
30-40	1.6	1.7	2.8	3.1	2.1	1.8	0.9	0.8	0.2	0.2
20-30	1.1	1.7	3.8	2.2	1.7	1.5	0.7	0.4	0.1	0.1
10-20	2.1	8.1	1.6	1.3	1	0.8	0.2	0.3	0.1	0.1
0-10	7.7	1.9	0.9	1.1	0.6	0.3	0.2	0.3	0.1	0.1
Total	**16.5**	**18.7**	**18.4**	**17.5**	**13.0**	**11.0**	**6.3**	**6.8**	**3.1**	**3.0**

Table 5: Global contact matrix related to EpIGraph's social model used in the experiments. Each column represents the average number of contacts of the individuals belonging to the age interval shown on the column's top. These contacts are broken down by age intervals in the different rows.

the group size in a random fashion. This means that the connection pattern of each group is unique, while certain graph-related properties like variable distribution of the number of contacts per individual [2] are maintained.

The social model includes two more types of social contacts for leisure and family activities. Leisure contacts are modelled by means of inter-group contacts. These contacts are between individuals belonging to different groups (for instance, work and school groups) and mostly occur after the main daily activity and before family time, as well as during the weekends. These contacts represent interactions with friends as well as casual contacts with unknown people. The third class of contacts are family contacts, interactions with family members who may or may not be part of the same group. The family connections graph is completely connected. However, connections are time-dependent which means that some of them (related to individuals that are not at home at a given time) are only active during a certain time interval (see Figure 1).

Recently, we have enhanced the social model combining the previous contact distribution with information of contact matrices. These matrices, extracted from surveys, contain statistical information of the average number of contacts between individuals of a certain ranges of ages. Table 4 shows the contact matrix used as input [5] in our social model that corresponds to Spain. Based on this table, individuals that are between 10 and 20 years old have an average of 2.1 and 8.3 contacts per day with other individuals in the range of ages of 0 to 10 years, and 10 to 20 years, respectively. This contact matrix is divided into four sub-matrices that differentiate between school, work, community and household contacts.

Based on these matrices, we have developed a new graph scaling algorithm that generates work connection graphs with a certain average connectivity $< k >$ based on the aforementioned contact matrices. The school sub-matrix was used to generate the student's school groups. The work sub-matrix was used to generate the workers' work groups and stay-at-home persons' and elders' informal meetup groups. Finally, the community sub-matrix was used to generate the leisure contacts reflecting the contacts that are not related neither to school nor work connections, for instance, child-worker, child-elderly or worker-elderly connections. Table 5 shows the resulting contact matrix that has been obtained from the contact model created by the simulator. Note that both of them (real and simulator-generated) have similar values.

Another new feature that EpiGraph includes are ad-hoc connections for specific collectives. These connections may be static or dynamic. Static connections are generated during the social model creation and are permanent during the simulation. Dynamic connections are generated during the simulation time according to the existing conditions and they may change over time. The following list describes the ad-hoc connections implemented in the simulator:

- Ad-hoc school connections are static. Each educator (worker belonging to the education sector) is in contact with all the students of a certain class during the work time slot.
- Ad-hoc elderly caregiver connections are also static. Each elderly caregiver is in contact with a certain group of elderly people at a nursing home during the work time slot.
- Ad-hoc health-care connections are dynamic. Each worker belonging to the health sector is in contact with 30 patients per day [2]. For non front-line health workers, the patients are chosen at random from the existing population, so the risk of meeting a COVID-19 infected individual is the same as for any other profession. In contrast, front-line health workers have 3.3 times more risk of meeting a COVID-19 infected patient.
- Ad-hoc catering connections are dynamic. Each worker belonging to this sector is in contact during the work time slot with 10 other individuals per hour. The contacts (that represent the customers) are chosen at random among the population individuals avoiding the ones that are in bed or hospitalized.
- Ad-hoc public security connections are dynamic. Each officer is in contact during the work time slot with 5 other individuals per hour. The contacts are chosen at random among the population individuals avoiding the ones that are in bed or hospitalized.

[2] This value corresponds to the daily average number of patients per doctor in Spain.

- Ad-hoc occasional meetings are also dynamic. They represents meetings between different groups related to social events. We consider three different classes of occasional meetings: (1) work occasional meetings represent meeting between people belonging to different work groups. Here, once per day we select a varying number of work groups (between 2 and 4) at random and we connect them during a four-hour time period. Note that only a small fraction of the existing work groups are connected by these procedure (the remaining ones do not participate on occasional meetings). (2) school occasional meetings are similar to the previous one but student's groups are chosen instead of work groups. They represent occasional social gathering between students. Finally, (3) leisure occasional meetings represent groups of friends that gather in social meetings. Like in the previous occasional meeting classes, these connections are created once per day (in this case, from the leisure contacts) from only a small fraction of the existing contacts.

Note that all these ad-hoc contacts are complementary to the existing individual's work contacts. In this way, a certain professional, during the work time will have to two different interaction: the graph-generated ones that connect the individual with other work colleagues (for instance, educators belonging to the same school or catering employees working at the same restaurant) and the ad-hoc ones that connects the individual with other contacts (educators with students and catering workers with customers).

2.3 Transportation model

The transportation model reflects the movement of people between cities for work, study, or vacation, and it is based on the gravity model proposed by Viboud et al. [7]. Note that the movement of people within a city is already captured by the social model. The transportation model serves the purpose of moving individuals between different cities, allowing for disease transmission over large areas. The geographical information that EpiGraph takes into account includes latitude, longitude, and distance between urban regions, and was extracted from the Google Maps web service using the Google Distance Matrix API [1].

2.4 COVID-19 model

The epidemic model implemented in EpiGraph is a compartmental stochastic SEIR model extended to include compartments for latent, asymptomatic, and dead, as well as hospitalized and vaccinated states. However, instead of being an analytic model based on differential equations, Epigraph follows an approach based on probabilities using randomness to determine the duration and transitions between the compartments. In addition, the basic reproduction numbers R_0s are different for each compartment. Figure 2 shows the infection phases, which are described below; The different infection stages are:

- **Incubation stage**. At the beginning of this stage individuals are infected but symptoms are not present and they are not yet able to transmit the virus. This stage is represented as primary exposed E^P. From this stage the infection can enter one of two phases, based on a probability P^{EI}: a secondary exposed stage E^S where slight symptoms appear and the individual becomes infectious with a certain R_0^{ES}, or an asymptomatic stage (described below).

- In the **asymptomatic stage** (compartment A), infected individuals do not notice symptoms but are able to transmit the disease with a certain R_0^A reproduction number. After a certain time, they pass to the recovered compartment in which the subject acquires viral immunity.

- In the first **symptomatic stage** - called primary infection state I^P - symptoms appear and a certain fraction of the individuals (given by a probability P^V) seek medical attention. This may imply initiating antiviral therapy (I_V^S state), which may reduce both the period of infectiousness and the individual ability to transmit the virus. Note that antivirals are only effective for a certain duration of time (called window of opportunity) from the beginning of the symptomatic stage. To what extent the antiviral treatment will have an effect depends on the time within the window when an individual seeks medical care. Instead of using a fixed duration for the window

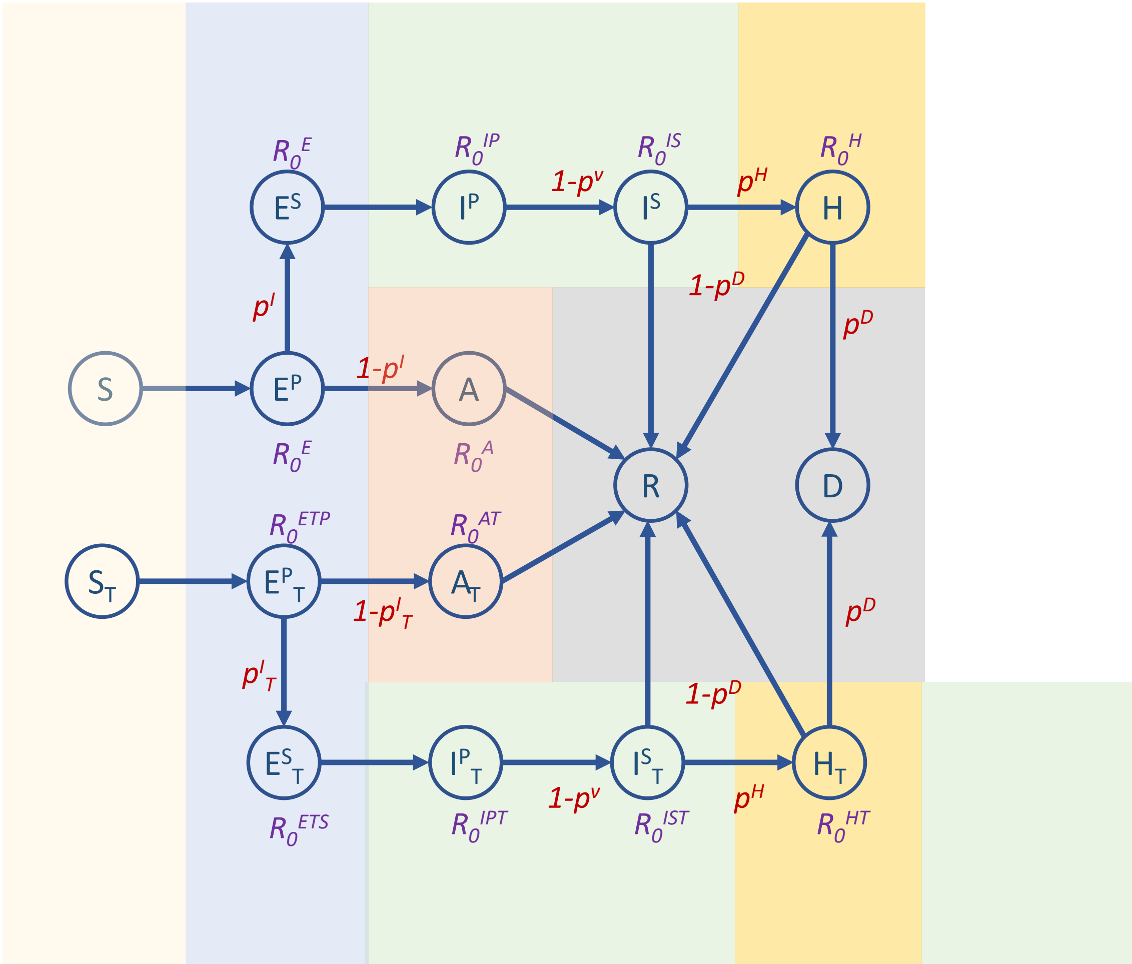


Figure 2: Compartmental model used by EpiGraph. It consists of the following states: susceptible (S), primary exposed (E^P), secondary exposed (E^S), asymptomatic (A), primary infectee (I^P), secondary infectee with antiviral treatment (I_V^S), hospitalized (H), recovered (R) and dead (D). Each state shows the basic reproduction number of the state (non existing R0s means that are not applicable). The edges show the transition probabilities (which are normalized) between the compartments. Duration of the main infection stages consists of an incubation that includes E^P and E^S; infectious includes I^P, I^S and I_V^S (which is not considered in our experiments); hospitalized is represented as H; and asymptomatic is A. Note that the asymptomatic stage starts after the primary exposed stage (E^P), which in this approximation lasts only one day. States S_T, E_T^P, E_T^S, A_T, I_T^P, I_T^S and H_T are related to *treated individuals* i.e. individual that have been vaccinated.

of opportunity, we assume that every individual may have a slightly different window size (by using a probability distribution). Individuals that do not receive this treatment during their window of opportunity will transition to phase I^S, where symptoms persist but antivirals are no longer effective. I^P, I^S and I^V have associated basic reproduction numbers of R_0^{IP}, R_0^{IS} and R_0^{IV}. In this work we assume that there are no available antiviral treatments.

- A certain fraction of the individuals are hospitalized (**hospitalized stage**). The probability of entering this stage is given by the parameter $P^H(age)$, which increases with age. From this state, an individual may transition to either the recovered or the dead stage. During hospitalization, we use R_0^H for modelling the transmission in hospitals.
- The individuals that reach the **dead stage** are removed from the simulation. The transition probability, denoted as $P^D(age)$, is also age-dependent and is applied over the portion of hospitalized individuals.
- The **treated stages** shown in the lower part of the figure represent the infection stages for vaccinated individuals. Section 2.5 describes these stages.

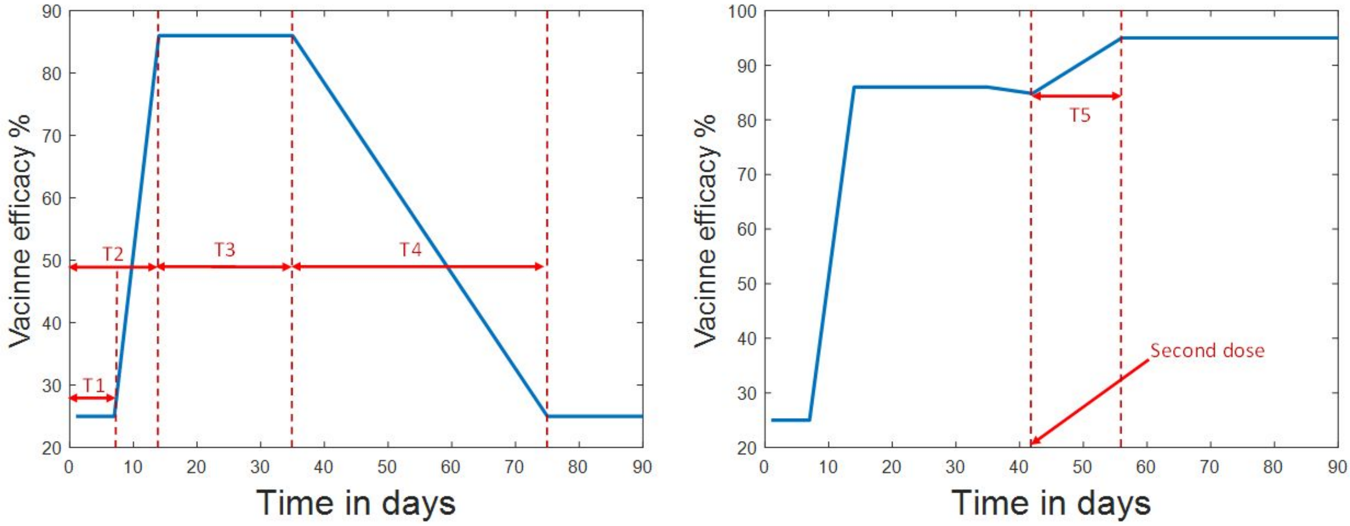


Figure 3: Generic COVID19 vaccine efficacy for the first dose (on the left) and second dose (on the right).

The time spent in a given state is generated following a normal distribution to simulate the time ranges specific to each stage of the infection and the fact that each individual may go through phases of different lengths. Figure 2 shows an overview of the different infection stages. We also consider that a percentage of the sick individuals stay in bed, thus reducing the number of people that they interact with.

2.5 Vaccination model

In this study a generic COVID19 vaccine was modelled by means of parameters as a use case. In the compartment model shown in Figure 2 a vaccinated individual that is infected (E_T^P state) transitions to two possible states: if the vaccine protects the individual, then the transition is to the Asymptomatic Treated state (A_T) and he/she will not suffer any health condition. Otherwise, when there is a vaccination failure, the transition is to the Exposed Secondary Treated (E_T^S) state. Then, it will transition to infected treated states I_T^P and I_T^S and the individual will be in risk of being hospitalized (H_T) or dead. We assume that the probabilities of developing health risks (probabilities of being hospitalized or death) are the same as for non-vaccinated individuals. The vaccination efficacy is modelled as the probability of transitioning to the A_T state. For instance, a vaccinated individual with an existing vaccine efficacy of 95% means that is infected will have a 95% of probability to transition to the A_T state.

Figure 3 (left) shows a generic COVID19 vaccine model for a single dose. The vaccine starts being effective after 7 days after the vaccination (T_1 parameter). Note that the background efficacy is 25%, that corresponds to the same probability for transitioning to the Asymptomatic state for a non-vaccinated individual. During the following 7 days ($T_2 - T_1$ parameters) the vaccine efficacy scales up to a certain value that was chosen 85%. If a second dose is not received, then, the first dose will keep this efficacy during a certain period (T_3 parameter) and then it linearly decreases until reaching the background value. The duration of this decreasing phase is given by T_4 parameter.

Figure 3 (right) illustrates the combined effect of two doses. In this example, the second dose is applied at day 42, when the vaccine's efficacy has begun to reduce. After that, it linearly increases to a maximum value that was chosen 95% in this example. The vaccine's increase time period is given by T_5 parameter.

Regarding the contagion model we distinguish three different transmission scenarios for the vaccinated individuals: *non contagious*, in which the vaccine prevents the contagion for all the states, *partially contagious* in which the contagiousness of the treated states is the same as the infected ones but the asymptomatic individuals do not transmit the disease, and *fully contagious with a probability P_C* where a vaccinated individual are contagious with a certain probability.

2.6 COVID-19 mitigation strategies

Given that the simulator considers every single individual and their connections, it is possible to model in detail the different social distancing and mitigation policies imposed by the authorities. We developed a new component called the *mitigation model* that links these policies with the social and the mobility models. The policies that we are considering are:

- Social distancing. EpiGraph distinguishes four classes of contacts between individuals: at school, at work, with family, and during leisure time. We leverage this distinction to evaluate social distancing policies that can apply differently to the contact types, for instance the closure of the schools and work places. We have considered both essential and non-essential workers - which represent 0.35% and 0.65% of company employees - as well as the interruption of leisure activities.
- Mobility restrictions. The transportation model of EpiGraph includes long and short-distance movements of individuals between cities. In this work we have introduced policies that restrict each one of them independently.
- Face-masks. We have introduced the use of surgical and ffp2-grade face-masks, which we evaluate when used by the general population or by targeted groups, such as the elderly.
- COVID-19 testing. In this mitigation strategy a given number of randomly-selected individuals are tested each day, starting on a given date. If the individual is infected, then several policies can be applied. For instance, only the individual is isolated at home, or the individual and all his/her related contacts are isolated as well.

References

[1] Google LLC, https://developers.google.com/maps/. *Google Maps API*, 2021.

[2] G. Martín, M.C. Marinescu, D.E. Singh, and J. Carretero. Leveraging social networks for understanding the evolution of epidemics. *BMC Syst Biol*, 5(S3), 2011.

[3] Gonzalo Martín, David E. Singh, Maria-Cristina Marinescu, and Jesús Carretero. Towards efficient large scale epidemiological simulations in epigraph. *Parallel Computing*, 42:88 – 102, 2015. Parallelism in Bioinformatics.

[4] Ministry of Economic Affairs and Digital Transformation (MINECO), http://www.ine.es/. *National Statistics Institute (INE)*, 2021.

[5] D. Mistry, M. Litvinova, A Pastore y Piontti, and et al. Inferring high-resolution human mixing patterns for disease modeling. *Nature Communications*, 12:323, 2021.

[6] David E. Singh, Maria-Cristina Marinescu, Miguel Guzmán-Merino, Christian Durán, Concepción Delgado-Sanz, Diana Gomez-Barroso, and Jesus Carretero. Simulation of covid-19 propagation scenarios in the madrid metropolitan area. *Frontiers in Public Health*, 9:172, 2021.

[7] Cécile Viboud, Ottar N Bjørnstad, David L Smith, Lone Simonsen, Mark A Miller, and Bryan T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451, 2006.