Received: 30 October 2020 | Revised: 17 May 2021 | Accepted: 5 June 2021

DOI: 10.1002/mma.7628

SPECIAL ISSUE PAPER

WILEY

Probabilistic calibration and short-term prediction of the prevalence herpes simplex type 2: A transmission dynamics modelling approach

Juan-Carlos Cortés | **Pablo Martínez-Rodríguez** | **José-Antonio Moraño** | **José-Vicente Romero** | **María-Dolores Roselló** | **Rafael-Jacinto Villanueva**

Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, Valencia, Spain

Correspondence
Pablo Martínez-Rodríguez, Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, Valencia, Spain.
Email: pabmarr2@upv.es

Communicated by: J. R. Torregrosa

Funding information
Generalitat Valenciana, Grant/Award Number: AICO/2019/215

An epidemiological model is proposed to study the transmission dynamics of the herpes virus type 2, a sexually transmitted infectious disease. This model considers two states, susceptible and infected, divides the population into sexes, assumes only heterosexual contacts and includes different transmission rates depending on whether the transmission is woman–man or man–woman. Reported and prevalence series data are retrieved from several sources. We consider the inherent data survey errors and the sensitivity of the diagnosis tests (data uncertainty). To calibrate the model to the available data and their uncertainty, a novel technique is proposed in two steps: (1) the application of the estimation of distribution algorithm (EDA) to find sets of model parameter values close to the data uncertainty and (2) the application of a selection algorithm to get a reduced number of model parameter values whose model outputs capture accurately the data uncertainty. Then, we check its robustness, and we provide a prediction of the evolution of the infected over the next 4 years. From the technical point of view, we conclude that the proposed technique to calibrate probabilistically the model is reliable and robust. Also, it is able to provide confidence intervals for the model parameter values and the predictions. From the medical point of view, the model returns that the transmission woman–man is higher than the man–woman, according to recent literature, and there is a mild increasing trend in the number of infected people over the next years.

KEYWORDS

computational technique, data uncertainty quantification, herpes simplex virus type 2, modelling infectious diseases, probabilistic calibration, transmission dynamics

MSC CLASSIFICATION

37H10; 39A50; 90-08; 90-08

1 | INTRODUCTION

Herpes simplex type 2 virus (HSV2) is a sexually transmitted disease. People infected with this virus sporadically suffer from painful blisters or ulcers at the area of infection along their life. Also, people who are infected with HSV2 significantly increase their possibility of contracting HIV infection.[1] In addition, pregnant women infected near the time of delivery may transmit the virus to her baby, producing on this baby a high risk to suffer from neurological disabilities or even death.[2,3] Furthermore, less relevant but more common is the fact that the presence of HSV2 generates social stigma and a negative psychosocial impact in the infected population.[4] Moreover, in the last times, a higher number of women with HSV2 have been reported,[5–7] which describes an upward trend in the infection transmission. All these facts described above define HSV2 infection as a concerning disease, and it motivates the interest to keep developing mathematical models to describe HSV2 dynamics in order to design efficient health care campaigns that help to reduce both HSV2 and, consequently, HIV new infections.

To date, various models have been proposed to analyse the dynamics of HSV2.[8–17] In Blower et al,[8] a model is proposed to study the effect of antiviral resistance on infection in a homogeneous population (without considering age and sex). The model considers regular latency periods where the infected subpopulation does not transmit the disease. In this study, values for model parameters are taken from literature. White and Garnett[9] also study the effect of antivirals considering a model that divides the population into two groups, depending on their level of sexual activity, and the model distinguishes between recently infected and those who have been infected for a while. In Schinazi,[10] a model to study the transmission dynamics of HSV2 is proposed, where homogeneous population and regular latency periods are considered. Newton and Kuder[11] proposed a model similar to Schinazi,[10] where the susceptible population goes through a period of exposure in a first stage before becoming infected. In Fisman et al,[12] a compartmental model is proposed where the evolution of the virus transmission and its economic cost are studied. The model divides the population into sexes and into three age groups, assuming only heterosexual relations and a constant contagion of the infected (no latent periods). In Korenromp et al,[13] a study is carried out on the impact of HSV2 on the transmission of HIV. The model divides the population into sexes, assuming only heterosexual contacts and defining increasing latency periods in the infection. Also, the model considers higher rates of infection when the infected are in early stages. In Garnett,[14] the impact of female partial vaccination on the transmission dynamics is analysed. The population is divided according to sex, assuming only heterosexual contacts and regular latency periods. Furthermore, the model distinguishes between asymptomatic and symptomatic infected individuals, assigning them different transmission rates. In Ghani and Aral,[15] a model is proposed where the involvement of sex workers and clients in the global transmission of the HSV2 population is studied. The model divides the population into sex workers and clients and analyses their contact with the general population. This study does not consider the latency periods of the disease. In Foss et al,[16] a complex model to study the evolution of the infection in a population made up of sexual service users is analysed. In this model, the population is divided into sexes, assuming only heterosexual infection and regular latency periods. In addition, it distinguishes the infected subpopulation into symptomatic and asymptomatic, as well as between recently infected and those who have been infected for a while, assigning different transmission rates to each of the states. However, the authors are forced to simplify the model to reduce the number of parameters. In Lou et al,[17] a study is conducted to analyse the impact on the transmission of the virus if young women were vaccinated. The compartmental model proposed in this study divides the population into sexes, stratifying the females into three age ranges and assuming only heterosexual infection. No latent periods are considered.

The main differences between the models previously indicated are the stratification of the infected subpopulation in its various states and the constant, periodic or decreasing nature of the rates of transmission in the dynamics of the disease.

Our approach to the problem consists of building a simple model dividing the population into sexes and considering only heterosexual intercourses. Then, we calibrate the model parameter values with updated infected data, retrieved from healthcare system reports, and taking into account their uncertainty. We have made this decision to keep a balance between the complexity that involves when dealing with the computational treatment of uncertainties in dynamical models and the available real data. Therefore, our approach focuses on using real data to calibrate the model taking into account the data and phenomenon uncertainty.

Comparing our model with the aforementioned papers in our bibliographic revision, in previous studies[9–12,14,15] have been performed theoretical dynamic analysis where prevalence data and model parameter values are not required. Previous studies[8,16,17] gather the model parameter intervals from meta-data analysis appeared in the literature and apply the Latin hypercube sampling (LHS)[18] to perform simulations that allow them to compute mean and confidence intervals (CIs). Neither prevalence data are used nor model calibration are performed. Korenromp et al[13] propose an agent-based

model using the software STDSIM that does not consider uncertainty in the calibration process and, consequently, only a sensitivity analysis is performed. Furthermore, the paper uses data from 1980–2000 in Africa.

Uncertainty is a key aspect in mathematical modelling, particularly in the context of epidemiology, since information utilized to set model inputs is mainly based upon surveys that may contain sampling errors. To this source of uncertainty, it must be added the partial knowledge of the disease transmission rates which, in mathematical modelling, are represented by parameters that embed complex factors as social behaviour or how the disease can be transmitted. This motivates that model parameters should be treated as random variables or stochastic processes rather than constants or deterministic functions, respectively.

We build a model that considers two states, susceptible and infected, and two populations, men and women. We have chosen this scheme because, as we mentioned before, the disease affects individuals depending on their sex.[5–7] In our model, we are going to introduce the uncertainty considering model parameters as random variables that we have to calibrate in order to provide reliable predictions. A critical point in this contribution is the development of a technique to estimate probabilistically the model parameter values.

We use real data from Catalonia (Spain) and calibrate the model parameters taking into account the data uncertainty. The employed data have been obtained from two different sources, reported cases and incidence and prevalence rates. Apart from these sources, we should also consider uncertainty coming from errors in the clinical analysis of the biological samples used for the diagnosis of the reported cases. For the probabilistic calibration, we design an ad hoc computational technique to find sets of model parameter values such that when substituting them into the model, the outputs capture the data uncertainty. The model parameter values found will help us to estimate the mean and the 95% CI of the model parameters. Afterwards, with the sets of model parameters obtained, we perform predictions of the prevalence of HSV2 over the next 4 years.

This probabilistic calibration process has been designed in two steps. In the first, we look for a 'sufficient' number model outputs around the data and their uncertainty. This first step has been carried out using the optimization algorithm estimation of distribution algorithm (EDA)[19] to guarantee the closeness between model outputs and data. This algorithm has been successfully applied in other continuous optimization problems.[20–22] In the second step, we select among the obtained model outputs those that fit, as much as possible, the data mean and the 95% CI using a selection algorithm inspired in the particle swarm optimization (PSO) algorithm.[23]

The joint use of these two algorithms is based on the fact that we are not performing a deterministic calibration. We know that the data contain errors, and we want to calculate a suitable probabilistically distribution for each model parameter so that all their model outputs capture the data and their errors. Therefore, we need to perform several EDAs calibration in such a way we can capture the data uncertainty. In other words, in the first step, we obtain a pile of model parameter values using EDA. We need a lot of close model parameter values encoded in probabilistic distributions obtained via kernel density estimation (KDE) to be sure that, in the second step, we can select an appropriate subset such that, as a whole, their model outputs capture accurately the data and their errors.

This paper is structured as follows. In Section 2, we explain the proposed mathematical model and the treatment we do with the available data. In Sections 3.1 and 3.2, we explain the EDA and the *PSO_S* calibration algorithms, respectively, and the obtained results. In Section 4, we demonstrate the robustness of the proposed process. In Section 5, we show the prediction. Finally, in Section 6, some conclusions are drawn.

2 | MODEL

2.1 | Data source and treatment

First of all, we want to point out the difficulty to find data about HSV2. In our opinion, this happens because of two reasons: first, the people's reticence to inform about personal sensitive information regarding sexual behaviour and, second, the low HSV2 incidence rate. Furthermore, the available data are scarce and have to be retrieved from different sources that are neither necessarily obtained in the same way nor format. In our opinion, this mathematical treatment adds uncertainty to data that should be considered.

After doing an extensive revision of the literature and in order to calibrate the model, we have found different data sources. The first sources correspond to data from Europe,[6,7] and we have employ them in a first step to define the initial infected population in Catalonia (prevalence) for year 2012. We will consider year 2012 as our initial condition ($t = 0$). The second set of sources corresponds to Catalonia (Spain) laboratory reports,[24–28] and we have employed them to calibrate the model along years 2012–2016.

From the source of Europe,[7] for population aged between 15 and 49, we have obtained the means and 95% CI about prevalence rates corresponding to year 2016, 0.06418 and [0.02949, 0.13356] for men, and 0.13009 and [0.06094, 0.26487] for women. Since only punctual information is available for year 2012 in Europe,[6] we assume the same proportions between the percentiles and the mean for years 2012 and 2016, obtaining the following means and intervals for year 2012, 0.05493 and [0.02524, 0.11431] for men and 0.12422 and [0.058192481, 0.252913477] for women.

We have assumed that the Catalonia's prevalence rate in 2012 is similar to Europe's in the same year and in the same range of the population. We multiply the rates considered for year 2012 by the population of Catalonia (Spain)[29] in the age range 15–49, obtained as the average during the period 2012–2016. This yields the central values and the intervals 104,567 and [48,044, 217,613] for men and 224,651 and [105,242, 457,398] for women.

The data employed for the calibration correspond to the reported cases of people aged 15–49 who were infected with HSV2 in Catalonia (Spain) in the period 2012–2016. The data series have been transformed into accumulated form, since in this way, the data become more regular to facilitate the subsequent model parameters calibration.

The available data (d_k^c, being $k=m$ for men and $k=w$ for women) represent the annual reported people and not the total infected populations. Therefore, in order to calibrate the HSV2 model, we have included two proportion parameters, $p_m(t)=\frac{R_m(t)}{\Delta I_m(t)}$ and $p_w(t)=\frac{R_w(t)}{\Delta I_w(t)}$, relating the reported cases ($R_k(t)$) to the new infected individuals ($\Delta I_k(t) = I_k(t)-I_k(t-1)$), being t the time. However, we cannot define the proportion functions, since we do not know the new infected population (incidence) at each time t. The only times where we know the new infected population are in years 2012 and 2016 (obtained from Looker et al[6] and James et al[7] with the same reasoning done with the initial population). At this point, we assume a linear function between the values corresponding to years 2012 and 2016 to infer the information for the period 2013–2015 and then constructing intervals, $p_w(t) = [p_w^l(t),\ p_w^u(t)]$ and $p_m(t) = [p_m^l(t),\ p_m^u(t)]$, for women and men, respectively, for the period 2012–2016. The linear functions defining these extremes of the intervals are

$$p_k^l(t) = \frac{p_k^l(48) - p_k^l(0)}{48}\ t + p_k^l(0),$$

$$p_k^u(t) = \frac{p_k^u(48) - p_k^u(0)}{48}\ t + p_k^u(0).$$

In our model, the time step t is measured in months. The time $t=0$ corresponds to year 2012 and $t=48$ to year 2016.

The lineal functions $p_k^l(t)$ and $p_k^u(t)$ have been obtained with the data shown in Table 1, where $\Delta I_k(t)$ has been obtained from Looker et al[6] and James et al[7] for years 2012 and 2016, respectively, and $R_k(t)$ has been obtained from previous reports[24,28] for years 2012 and 2016, respectively.

From the reported data, it has been assumed an error interval of 15% variation with respect to the central value (d_k^c), since the source of the data corresponds to the tests carried out in laboratories. These tests have different sensitivity and reliability that depend on the nature of the test. As it is exposed in Navarro Ortega et al,[30] the best test has 85% sensitivity. Then, in order to capture the uncertainty in reported data, we build intervals centred at d_k^c as $[d_k^l, d_k^u] := [0.85d_k^c, 1.15d_k^c]$. The data are shown in Table 2.

TABLE 1 Proportion parameters ($p_k(t)$) of men and women in Catalonia (Spain) in 2012 and 2016

		Lower 95% CI limit			Upper 95% CI limit		
		$\Delta I_k(t)$	$R_k(t)$	$p_k^l(t)$	$\Delta I_k(t)$	$R_k(t)$	$p_k^u(t)$
$t=2012$	$k=w$	4659	99	0.0213	18,635	99	0.0053
	$k=m$	2156	81	0.0373	11,858	81	0.0068
$t=2016$	$k=w$	4985	291	0.0584	19,941	291	0.0146
	$k=m$	2008	144	0.0717	11,044	144	0.0130

TABLE 2 Women and men aged 15–49 reported with HSV2 from Catalonia in the period 2012–2017 and their allocated intervals

	2012	2013	2014	2015	2016
d_w^u	114	286	545	842	1177
d_w^c	99	249	474	732	1024
d_w^l	84	212	403	623	870
d_m^u	93	224	390	560	726
d_m^c	81	195	339	487	631
d_m^l	68	165	288	414	536

2.2 | Model building

Mathematical models are useful tools that help to understand the transmission dynamics of infectious diseases. The procedure of this work consists in developing an epidemic model of the HSV2 in Catalonia (Spain) along 2012–2016 for people aged 15–49. Taking into account that HSV2 is a chronic sexually transmitted illness and the availability of data from official sources, we are going to only consider two states, susceptible and infected, for women and men. Then, we define the following subpopulations:

- $S_w(t)$ denotes the number of susceptible women aged 15–49 at month t.
- $S_m(t)$ denotes the number of susceptible men aged 15–49 at month t.
- $I_w(t)$ denotes the number of infected/infectious women aged 15–49 at month t.
- $I_m(t)$ denotes the number of infected/infectious men aged 15–49 at month t.

Now, $S_w(t)+I_w(t)=P_w(t)$ is the total women population. Analogously, $S_m(t)+I_m(t)=P_m(t)$ is the total men population. A woman enters in the system (susceptible state) at rate μ_w, the same for men at rate μ_m. The entry into the system is modelled for women and men by the terms

$$\mu_w P_w(t) \quad \text{and} \quad \mu_m P_m(t),$$

respectively. Taking into account the low mortality rate of young people in Catalonia (Spain), we can assume that μ_w and μ_m are the birth rates.

Women and men leave the system dying out or turning 50 at rates δ_w and δ_m, respectively. Infection increases the death rate in the neonatal population,[3] but as our population is aged 15–49, we consider that infection does not increase death rate. Then, women and men come out the system by the terms

$$\delta_w S_w(t), \delta_w I_w(t) \quad \text{and} \quad \delta_m S_m(t), \delta_m I_m(t),$$

respectively.

We assume the classical hypothesis of population homogeneous mixing for heterosexual intercourses,[31,32] that is, any infectious man (woman) may contact and infect any susceptible woman (man). Although it is known that infection transmission is more likely when the infected individual has an episode of ulcers, we assume a constant transmission rate between the susceptible and infected populations over time, since we have not specific information about this fact in order to consider it in the modelling formulation.

As previously indicated, in this model, we only consider that the sexual contacts are only heterosexual. From the prevalence data previously shown, we can appreciate different infection rates between men and women. Hence, we consider two different transmission rates, β_1 and β_2. Thus, a susceptible woman (man) may get infected of HSV2 by sexual contacts with infected men (women), and this contagion can be modelled by the non-linear term for women and men

$$\beta_1 S_w(t)\frac{I_m(t)}{P_m(t)} \quad \text{and} \quad \beta_2 S_m(t)\frac{I_w(t)}{P_w(t)},$$

respectively.

This model is mathematically defined by the following system of non-linear difference equations (1), where t is the time in months and β_1, β_2, δ_w, and δ_m are the model parameters:

$$S_w(t+1) = S_w(t) + \mu_w P_w(t) - \delta_w S_w(t) - \beta_1 S_w(t)\frac{I_m(t)}{P_m(t)}, \tag{1a}$$

$$I_w(t+1) = I_w(t) + \beta_1 S_w(t)\frac{I_m(t)}{P_m(t)} - \delta_w I_w(t), \tag{1b}$$

$$S_m(t+1) = S_m(t) + \mu_m P_m(t) - \delta_m S_m(t) - \beta_2 S_m(t)\frac{I_w(t)}{P_w(t)}, \tag{1c}$$

$$I_m(t+1) = I_m(t) + \beta_2 S_m(t)\frac{I_w(t)}{P_w(t)} - \delta_m I_m(t). \tag{1d}$$

The model dynamics can be described graphically as in Figure 1.

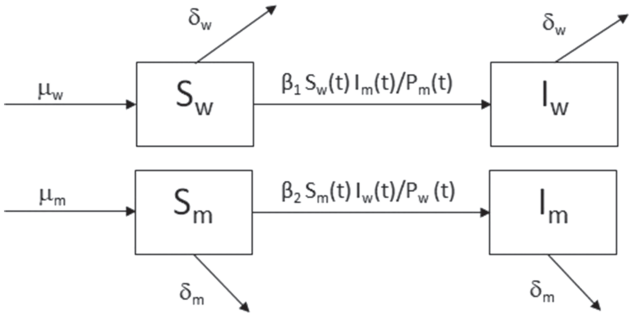


FIGURE 1 Epidemic model scheme for women and men corresponding to the system of non-linear difference equations (1a–1d)

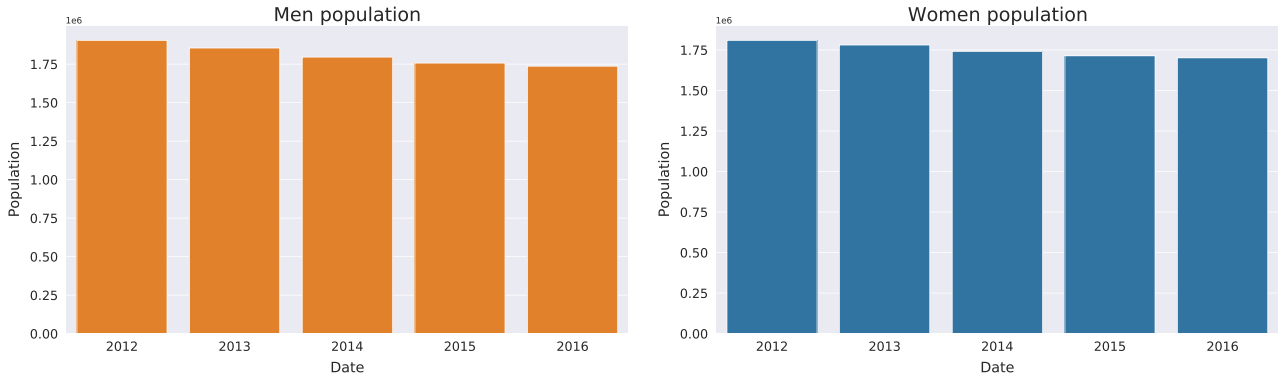


FIGURE 2 Population aged 15–49 in Catalonia (Spain) in the period 2012–2016 [Colour figure can be viewed at wileyonlinelibrary.com]

Furthermore, in our model, women and men populations are going to be considered as constant, since this way, the model flow in and flow out rates do not depend on time. This assumption can be also corroborated in Figure 2, where the population is almost constant. This fact entails the equations

$$P_w(t) = P_w = S_w(t) + I_w(t),\ P_m(t) = P_m = S_m(t) + I_m(t),\ \mu_w = \delta_w,\ \mu_m = \delta_m.$$

Here, in order to be consistent with the constant population hypothesis, we have two possibilities. We can consider the average birth rate from the demographic data, or we can consider the average death rate plus the rate of the exit of the system. In both cases, the results are very similar. Therefore, we have taken $\mu_w = \delta_w = 0.00286$ for women and $\mu_m = \delta_m = 0.00325$ for men, where the death rates plus the rate of the exit of the system have been obtained from National Statistics Institute.[29]

Assuming these simplifications and taking into account that $S_w(t) = P_w - I_w(t)$ and $S_m(t) = P_m - I_m(t)$, the new model is defined by the following system of non-linear difference equations (2):

$$I_w(t+1) = I_w(t) + \beta_1(P_w - I_w(t))\frac{I_m(t)}{P_m} - \delta_w I_w(t), \tag{2a}$$

$$I_m(t+1) = I_m(t) + \beta_2(P_m - I_m(t))\frac{I_w(t)}{P_w} - \delta_m I_m(t). \tag{2b}$$

3 | MODEL CALIBRATION

In this section, we are going to calibrate the model parameters with the available data and their uncertainty as described in the Section 2.1. To do that, we are going to propose a procedure split into two steps: (1) the application of the EDA to find model parameter values whose outputs lie inside or close to the CIs determining the data uncertainty and (2) the application of a selection algorithm based on classical PSO algorithm returning a set of model parameter values whose outputs capture, as much as possible, the data uncertainty.

3.1 | Step 1: Obtaining sets of model parameter values close to the data uncertainty

EDA[19] is a heuristic optimization algorithm that learns and samples a new generation of particles (model parameter values to be calibrated) from a probability distribution defined by the best particles in the previous generation.

In our case, the model parameters to be calibrated are β_1 and β_2, that is, the transmission rates men to women and women to men, respectively.

The objective of Step 1 is the generation of a repository formed by all the solutions (β_1, β_2), their model outputs and their fitnesses, calculated along the evolution of the algorithm such that a large number of them will be close to the data uncertainty.

The model has been simulated with t in months, and after the simulation, we have transformed the infected people into reported people $(R_w(t), R_m(t))$, applying the proportion parameters $p_w(t)$ and $p_m(t)$, for women and men, respectively. The proportion parameters have been obtained from a uniform sample within their ranges. Besides, as we need for the calibration the reported values in years, we have transformed them into annual periods, being now $t = 0, \dots, 4$, and corresponding to years 2012, $\dots$, 2016.

For this purpose, the fitness function $F(\beta_1, \beta_2)$ to be minimized by EDA is defined as follows:

1. Substitute the model parameter values (β_1, β_2) into the model, run a simulation and obtain $\{R_w(t), R_m(t)\}_{t=0}^{4}$, where $t = 0, \dots, 4$, corresponding to years 2012, $\dots$, 2016, the model outputs for women and men, respectively.
2. Calculate the distance between the obtained model outputs with the data CIs as follows:

$$D = \sum_{t=0}^{4} d\left(R_w(t), [d_w^l(t), d_w^u(t)]\right) + d\left(R_m(t), [d_m^l(t), d_m^u(t)]\right),$$

where the distance d from a point to an interval is defined by

$$d(x, [a, b]) = \begin{cases} 0 & \text{if} \quad x \in [a, b], \\ \min\left(|x - a|, |x - b|\right) & \text{if} \quad x \notin [a, b]. \end{cases}$$

Notice that this fitness function $F\colon\ ]0, \infty[\times]0, \infty[\to]0, \infty[$ takes two model parameter values and measures the closeness of their model output to data uncertainty. We want to minimize the function F. To do so, we use the following adaptation of the optimization algorithm EDA:

1. Input:
   - N, number of solutions in each generation.
   - M, number of solutions to be eliminated in every generation.
   - E, number of elite solutions.
   - G, number of generations.
2. Initialization: Generation of (β_1^i, β_2^i), $i = 1, \dots, N$ model parameter values, randomly.
3. Do G times:
   (a) Calculate the fitnesses $D^i = F(\beta_1^i, \beta_2^i)$, $i = 1, \dots, N$.
   (b) Sort the solutions by their fitnesses.
   (c) Eliminate M solutions with the worst fitnesses.
   (d) With the $N - M$ non eliminated solutions, we build a computational distribution K using an algorithm for KDE.
   (e) Eliminate all the solutions except the E solutions with best fitnesses (elite solutions).
   (f) Sample $N - E$ solutions from the previous calculated KDE K.
   (g) Join the $N - E$ solutions just sampled to the E elite solutions.

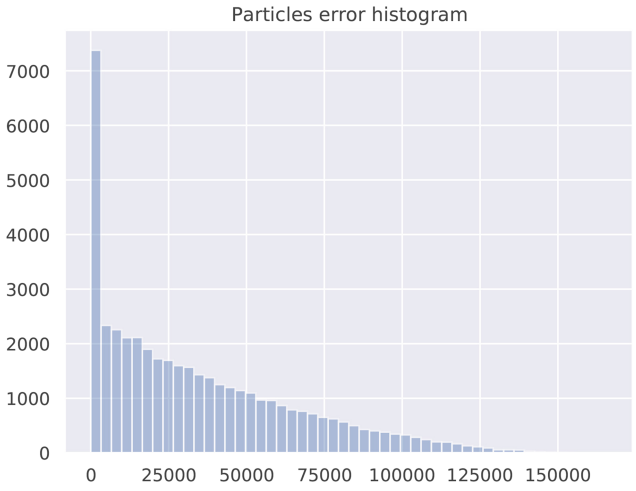
The probability distribution generated in each iteration in Step (e) is a non-parametric multivariate Gaussian kernel distribution (KDE).[33]

In order to facilitate the convergence of the EDA, it has been considered the so called *elite solutions*, in order to keep the best solutions from one generation to the next one.

TABLE 3 Repository structure

Index	Parameters	Model output	Fitness
1	(β_1^1, β_2^1)	$(\{R_w^1(t), R_m^1(t)\}_{t=0}^4)$	$D^1 = F(\beta_1^1, \beta_2^1)$
⋮	⋮	⋮	⋮
H	(β_1^s, β_2^s)	$(\{R_w^s(t), R_m^s(t)\}_{t=0}^4)$	$D^s = F(\beta_1^s, \beta_2^s)$

FIGURE 3 Histogram of the 50,000 solutions' fitnesses obtained [Colour figure can be viewed at wileyonlinelibrary.com]



During the execution of the EDA, we store the solutions in a repository, in such a way that, when the algorithm finishes, the solutions (β_1^s, β_2^s), their model outputs $(\{R_w^s(t), R_m^s(t)\}_{t=0}^4)$ and their fitnesses $(F(\beta_1^s, \beta_2^s))$, $s = 1, \ldots, H$, $H < N \times G$, are stored. The repository is structured as shown in Table 3.

For the implementation of the algorithm, we have used Python3 programming language and the Numpy[34] and Scipy[35] packages.

The parameters calibrated with the algorithm are the transmission rates β_1 and β_2. In order to facilitate calibration, the following search domains have been specified:

$$\beta_1 \in [0, 0.065], \beta_2 \in [0, 0.065].$$

These domains have been chosen after running some times the calibration and observing the convergence region of the model.

We have performed several runs of the EDA algorithm with different configuration parameters. The best results have been obtained with $G = 500$ generations and a population of $N = 100$ solutions per generation; in each generation, we have eliminated $M = 50$ solutions, and we have selected $E = 10$ elite solutions. Then, the total amount of evaluations of the model has been 50,000 (repeating the elite solutions). In Figure 3, we show the histogram of the 50,000 solutions' fitnesses obtained. As we can observe, most of the fitnesses are around 0, satisfying our goal of using an optimization algorithm to obtain a large number of model parameter values whose output is close to the data uncertainty.

3.2 | Step 2: Selection of the solutions with the *PSO_S* algorithm

Step 1 has provided a repository whose structure can be seen in Table 3. Then, we apply a selection algorithm inspired on PSO, called PSO selection (*PSO_S*).[36] At this point, let us suppose that the repository is made up of H solutions. The objective of the process is the selection of a reduced set of solutions from the repository, whose model output reproduces, as much accurate as possible, the CI of the data.

In order to measure the accuracy of a set of solutions, we are going to refer to the solutions by their index in the repository. Thus, if we have a set of indices $I = \{i_1, \ldots i_n\}$, we take the model outputs corresponding to the indices in I, for women:

Index	Model outputs				
i_1	$R_w^{i1}(0)$	$R_w^{i1}(1)$	$R_w^{i1}(2)$	$R_w^{i1}(3)$	$R_w^{i1}(4)$
…	…	…	…	…	…
i_n	$R_w^{in}(0)$	$R_w^{in}(1)$	$R_w^{in}(2)$	$R_w^{in}(3)$	$R_w^{in}(4)$
Perc. 2.5	$P_w^l(0)$	$P_w^l(1)$	$P_w^l(2)$	$P_w^l(3)$	$P_w^l(4)$
Mean	$m_w(0)$	$m_w(1)$	$m_w(2)$	$m_w(3)$	$m_w(4)$
Perc. 97.5	$P_w^u(0)$	$P_w^u(1)$	$P_w^u(2)$	$P_w^u(3)$	$P_w^u(4)$

and for men:

Index	Model outputs				
i_1	$R_m^{i1}(0)$	$R_m^{i1}(1)$	$R_m^{i1}(2)$	$R_m^{i1}(3)$	$R_m^{i1}(4)$
…	…	…	…	…	…
i_n	$R_m^{in}(0)$	$R_m^{in}(1)$	$R_m^{in}(2)$	$R_m^{in}(3)$	$R_m^{in}(4)$
Perc. 2.5	$P_m^l(0)$	$P_m^l(1)$	$P_m^l(2)$	$P_m^l(3)$	$P_m^l(4)$
Mean	$m_m(0)$	$m_m(1)$	$m_m(2)$	$m_m(3)$	$m_m(4)$
Perc. 97.5	$P_m^u(0)$	$P_m^u(1)$	$P_m^u(2)$	$P_m^u(3)$	$P_m^u(4)$

Perc. denotes the percentile function.

For each column in the above table, we calculate the percentiles 2.5, 97.5 and the mean. These values are compared with the percentiles 2.5, 97.5 and the mean of the data $\{d_w^l(t), d_w^c(t), d_w^u(t)\}$, $\{d_m^l(t), d_m^c(t), d_m^u(t)\}$, $t = 0, \ldots, 4$ (see Table 2) using the SMAPE measure:[37]

$$\begin{aligned} SMAPE(I) = \frac{100\%}{5}\sum_{t=0}^{4}\Bigg(& \frac{|d_w^l(t) - P_w^u(t)|}{|d_w^l(t)| + |P_w^u(t)|} \\ & + \frac{|d_w^c(t) - m_w(t)|}{|d_w^c(t)| + |m_w(t)|} + \frac{|d_w^u(t) - P_w^l(t)|}{|d_w^u(t)| + |P_w^l(t)|} \\ & + \frac{|d_m^l(t) - P_m^u(t)|}{|d_m^l(t)| + |P_m^u(t)|} + \frac{|d_m^c(t) - m_m(t)|}{|d_m^c(t)| + |m_m(t)|} \\ & + \frac{|d_m^u(t) - P_m^l(t)|}{|d_m^u(t)| + |P_m^l(t)|}\Bigg). \end{aligned} \tag{3}$$

Once defined the fitness function that measures the accuracy of a set of solutions to data uncertainty, we introduce the *PSO_S* algorithm to find the best set of particles (solutions) as follows:

1. Input:
   - N, number of particles in each generation of *PSO_S*.
   - n, number of indices in each particle.
   - *ITMAX*, maximum number of iterations.
   - α, percentage of new random indices in each generation.
2. Define $I_{global}^{best} = \varnothing$ and $SMAPE(I_{global}^{best}) = +\infty$.
3. Do N times:
   (a) Initialize I_i with a set of n indices chosen randomly without repetitions.
   (b) Evaluate its fitness $SMAPE(I_i)$.
   (c) Define its individual best fitness as $I_i^{best} = I_i$.
   (d) If $SMAPE(I_i) < SMAPE(I_{global}^{best})$ then $I_{global}^{best} = I_i$.
4. Do *ITMAX* times:
   (a) Do N times:

(i) Build the new set $P = I_i \cup I_i^{best} \cup I_{global} \cup I_\alpha$, that is, joining the current particle, its individual best, the global best and a random set of $\alpha \times N$ indices.
(ii) Remove the repeated elements.
(iii) Build the new particle I_i as the random selection without repetition of n elements of P.
(iv) If $SMAPE(I_i) < SMAPE(I_i^{best})$, then $I_i^{best} = I_i$.
(v) If $SMAPE(I_i) < SMAPE(I_{global}^{best})$, then $I_{global}^{best} = I_i$.

In order to streamline and facilitate the calibration of the *PSO_S* algorithm, a preliminary screening of the repository has been carried out. With this screening, those solutions whose fitnesses are greater than a given threshold have not been considered, since we are only interested in those solutions that are close to the CI of the data.

The number of solutions has been selected after running the screening with five different configurations. We have run the screening with 1000, 2000, 3000, 4000 and 5000 solutions. We have chosen these values in order to ensure that we select solutions near the CI of the data but not all of them inside their ranges because, eventually, we have to calculate the 95% CI. After performing all the configurations, the best result returned by *PSO_S* has been obtained with 5000 solutions.

If we compare Figures 3 and 4, we can observe the result of the screening process. The result of this process, in terms of the model output, can be observed in Figure 5, where the solutions of the screened repository are allocated around the CI of the data, for men and women. With this step, we have achieved a better *PSO_S* input.

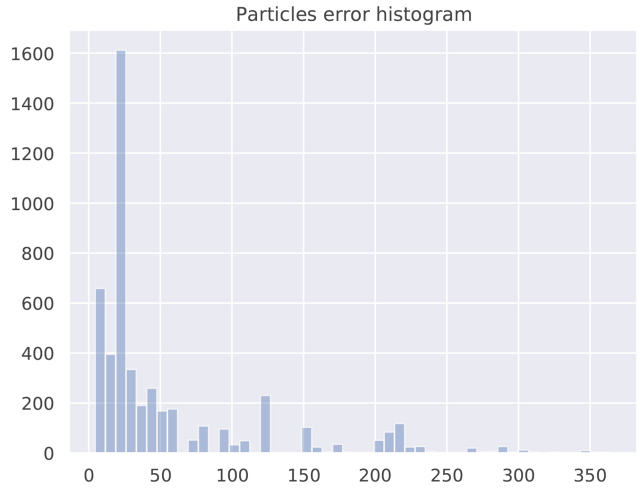


FIGURE 4 Histogram of the 5000 solutions' fitnesses screened that provide, after applying *PSOS_S*, the best result [Colour figure can be viewed at wileyonlinelibrary.com]

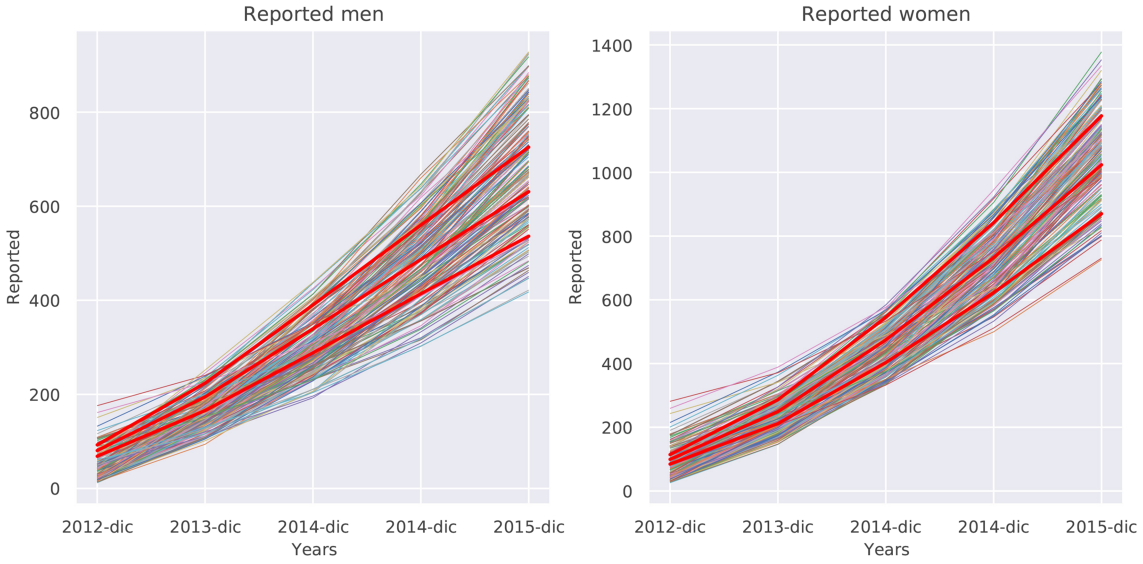


FIGURE 5 This figure shows how the screening process keeps model outputs (coloured lines) around the data uncertainty (red lines) discarding those with error greater than a given threshold [Colour figure can be viewed at wileyonlinelibrary.com]

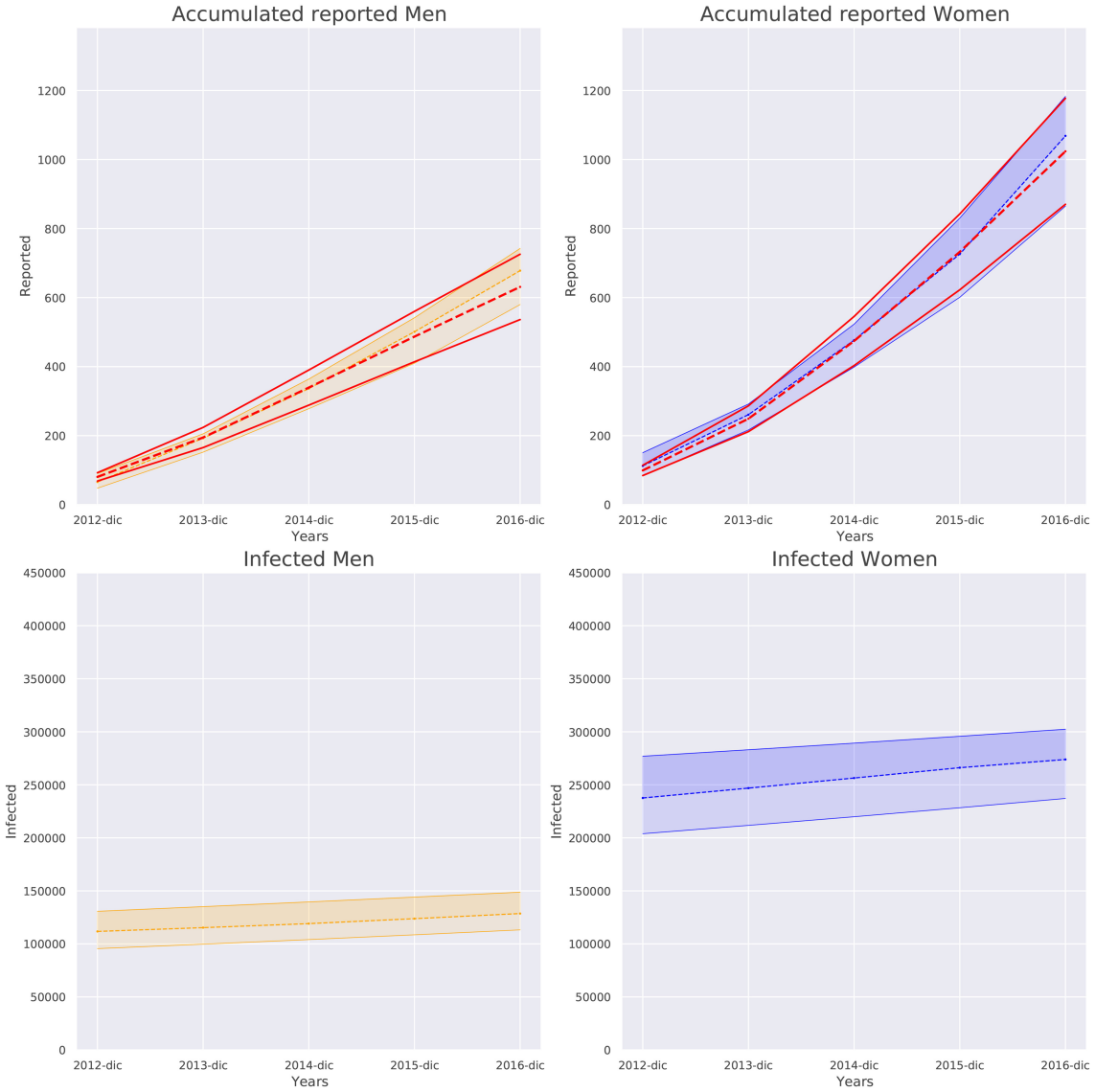


FIGURE 6 Graphical assessment of the probabilistic calibration. The upper figures show the accumulated data. The red lines are the mean and 95% CI of data, and the shadowed area is the calibrated model output. In the lower figures, we can se the number of infected given by the calibrated output [Colour figure can be viewed at wileyonlinelibrary.com]

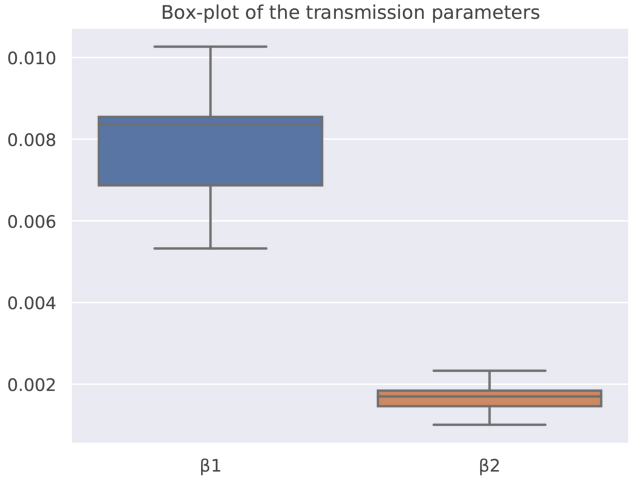


FIGURE 7 Box plot of the probabilistic model parameters [Colour figure can be viewed at wileyonlinelibrary.com]

We have performed a run of the *PSO_S* algorithm configured with $ITMAX = 30{,}000$ generations and a population of $N = 60$ particles per generation. Each one of this particles are formed by $n = 50$ indices referencing the solutions of the repository. The randomness parameter has been defined as $\alpha = 0.2$. The result of the process is a set of $n = 50$ EDA's solutions whose model outputs capture the data uncertainty with 7.80% of *SMAPE* error. This CI is shown in Figure 6.

TABLE 4 Study of the robustness of the proposed technique

		β_1		β_2	
Execution	**Error**	**Mean**	**95% CI**	**Mean**	**95% CI**
1	7.40%	0.00776	[0.00532, 0.00981]	0.00165	[0.00112, 0.00205]
2	10.09%	0.00754	[0.00558, 0.01155]	0.00155	[0.00120, 0.00208]
3	8.38%	0.00812	[0.00593, 0.00998]	0.00166	[0.00119, 0.00212]
4	8.85%	0.00765	[0.00576, 0.01200]	0.00164	[0.00117, 0.00233]
5	7.67%	0.00849	[0.00545, 0.01082]	0.00179	[0.00116, 0.00220]

Note: Errors and 95% CI of the model parameters are very similar.

TABLE 5 Women and men reported with HSV2 from Catalonia in 2017 and their defined confidence interval[38]

	d_w^u	d_w^c	d_w^l	d_m^u	d_m^c	d_m^l
2017	1620	1409	1197	947	824	700

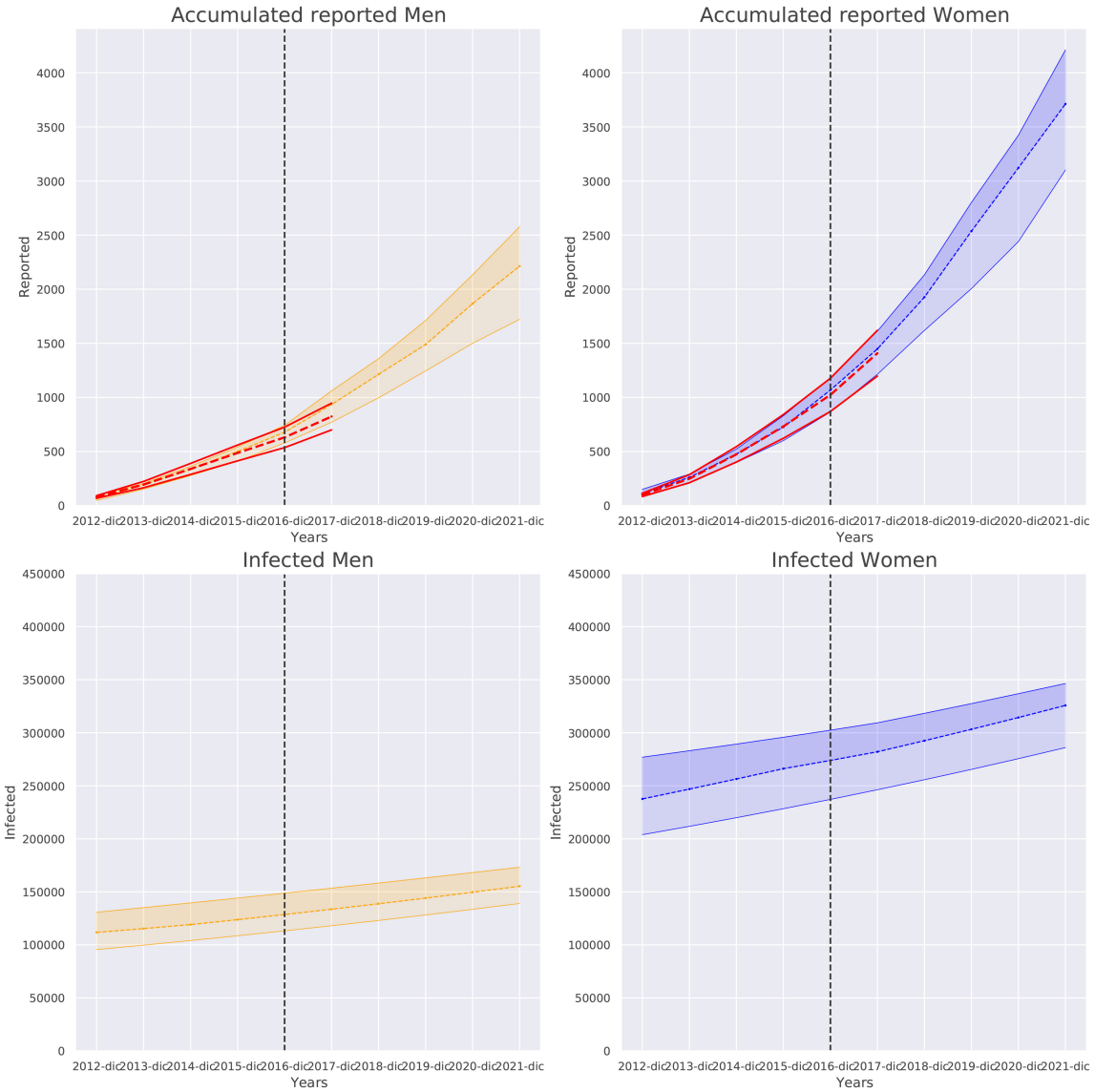


FIGURE 8 Prediction of the number of infected, men and women, over the next 4 years. The dotted black line shows the end of 2016. In the upper figures, we have the accumulated reported. We can see that our model has predicted very accurately the data reported in 2017. In the lower figures, we can see the predicted number of women and men infected [Colour figure can be viewed at wileyonlinelibrary.com]

The CIs and the mean of the $n = 50$ pairs of β_1, β_2 parameter values defining the transmission rates are 0.00776 and [0.00532, 0.00981] for β_1 and 0.00165 and [0.00112, 0.00205] for β_2. Also, we can see a graphical representation in Figure 7. The domain of the parameter β_1 is greater than the one for β_2, in clear agreement with publications that expose that transmission from male to female is greater than from female to male.[5–7]

For the implementation of the algorithm, we have also used Python 3 programming language and the Numpy[34] and Scipy[35] packages.

4 | ROBUSTNESS

In order to ensure that the proposed technique is robust, we repeated the procedure four times more. The results previously shown in Section 3 correspond to the first run. Table 4 shows the 95% CI of the model parameter values obtained after the runs and their *SMAPE* errors. As it can be observed, all the runs returned similar results and similar errors.

5 | PREDICTION

Once we have calibrated the model and we have proved that the obtained result is robust, we have run the model until 2021, and we have compared the CI of the model output with the CI of the data from 2017,[38] recently published after the realization of this work and shown in Tablec 5. As it can be observed in Figure 8, the model captures the CI of the new data from 2017 and provides a prediction over the next years.

6 | CONCLUSION

In this paper, we have proposed a system of non-linear difference equations to model the transmission dynamics of HSV2 in the Catalonia population along years 2012–2016. Also, we have taken into account the uncertainty inherent in the data and the fact that the available data correspond to reported cases and not to infected ones.

The technique proposed to calibrate the model represents a good strategy in order to capture the data uncertainty as far as the results reproduce in an acceptable way the uncertainty of the data (Figure 6). The transmission parameters obtained (Figure 7) confirm the fact that the transmission is higher from men to women than from women to men, as it is exposed in previous studies.[5–7] Furthermore, we can consider that the result obtained is robust as we have shown in Section 4 and with the validation in 2017 of the prediction in Section 5.

The use of EDAs has proven to be an appropriate tool for the calibration of the parameters in epidemiological models. In this case, it has been applied with the *PSO_S* algorithm. Regarding the proposed model, in the future, we will consider expanding the model to more complex structures where the latency periods will be taken into account.

The obtained results lead us to conclude two facts. The first one is that the Catalonia healthcare system is improving their ability to report cases if we assume as real the trend of Europe, and the second one is that as Europe, Catalonia has suffered a mild increase in the infected population of the virus between years 2012 and 2016 and that the trend for future years is similar.

ACKNOWLEDGEMENTS

This work has been partially supported by the Generalitat Valenciana (Grant AICO/2019/215). Authors are fully grateful for the comments suggested by the two reviewers.

CONFLICT OF INTEREST

This work does not have any conflicts of interest.

ORCID

Pablo Martínez-Rodríguez https://orcid.org/0000-0002-6878-1646
José-Antonio Moraño https://orcid.org/0000-0003-4385-7277
José-Vicente Romero https://orcid.org/0000-0003-3366-6557
Rafael-Jacinto Villanueva https://orcid.org/0000-0002-0131-0532

REFERENCES

1. Looker KJ, Elmes JAR, Gottlieb SL, et al. Effect of HSV-2 infection on subsequent HIV acquisition: an updated systematic review and meta-analysis. *Lancet Infect Diseases*. 2017;17(12):1303-1316. https://doi.org/10.1016/s1473-3099(17)30405-x
2. Brown ZA, Selke S, Zeh J, et al. The acquisition of herpes simplex virus during pregnancy. *New England J Med*. 1997;337(8):509-516. https://doi.org/10.1056/nejm199708213370801
3. Pinninti S, Kimberlin D. Maternal and neonatal herpes simplex virus infections. *Am J Perinatology*. 2013;30(2):113-120. https://doi.org/10.1055/s-0032-1332802
4. Cunningham NC, Zimet GD, Aalsma MC, Bernstein DI, Rosenthal SM. 35: expressed intent and acceptance of HSV-2 testing in adolescents. *J Adolescent Health*. 2006;38(2):130-131. https://doi.org/10.1016/j.jadohealth.2005.11.111
5. Looker K. An estimate of the global prevalence and incidence of herpes simplex virus type 2 infection. *Bull World Health Organ*. 2008;86(10):805-812. https://doi.org/10.2471/blt.07.046128
6. Looker KJ, Magaret AS, Turner KME, Vickerman P, Gottlieb SL, Newman LM. Global estimates of prevalent and incident herpes simplex virus type 2 infections in 2012. *PLoS ONE*. 2015;10(1):e114989. https://doi.org/10.1371/journal.pone.0114989
7. James C, Harfouche M, Welton NJ, et al. Herpes simplex virus: global infection prevalence and incidence estimates, 2016. *Bull World Health Organ*. 2020;98(5):315-329. https://doi.org/10.2471/blt.19.237149
8. Blower SM, Porco TC, Darby G. Predicting and preventing the emergence of antiviral drug resistance in HSV-2. *Nat Med*. 1998;4(6):673-678. https://doi.org/10.1038/nm0698-673
9. White PJ, Garnett GP. Use of antiviral treatment and prophylaxis is unlikely to have a major impact on the prevalence of herpes simplex virus type 2. *Sex Transm Infect*. 1999;75(1):49-54. https://doi.org/10.1136/sti.75.1.49
10. Schinazi RB. Strategies to control the genital herpes epidemic. *Math Biosci*. 1999;159(2):113-121. https://doi.org/10.1016/s0025-5564(99)00024-3
11. Newton EAC, Kuder JM. A model of the transmission and control of genital herpes. *Sex Transm Dis*. 2000;27(7):363-370. https://doi.org/10.1097/00007435-200008000-00001
12. Fisman DN, Lipsitch M, Hook EW, Goldie SJ. Projection of the future dimensions and costs of the genital herpes simplex type 2 epidemic in the United States. *Sex Transm Dis*. 2002;29(10):608-622. https://doi.org/10.1097/00007435-200210000-00008
13. Korenromp EL, Bakker R, Vlas SJD, Robinson NJ, Hayes R, Habbema JDF. Can behavior change explain increases in the proportion of genital ulcers attributable to herpes in Sub-Saharan Africa?. *Sex Transm Dis*. 2002;29(4):228-238. https://doi.org/10.1097/00007435-200204000-00008
14. Garnett GP. The potential epidemiological impact of a genital herpes vaccine for women. *Sex Transm Infect*. 2004;80(1):24-29.https://doi.org/10.1136/sti.2002.003848
15. Ghani AC, Aral SO. Patterns of sex worker–client contacts and their implications for the persistence of sexually transmitted infections. *J Infect Dis*. 2005;191(s1):S34-S41. https://doi.org/10.1086/425276
16. Foss AM, Vickerman PT, Chalabi Z, Mayaud P, Alary M, Watts CH. Dynamic modeling of herpes simplex virus type-2 (HSV-2) transmission: issues in structural uncertainty. *Bull Math Biol*. 2009;71(3):720-749. https://doi.org/10.1007/s11538-008-9379-1
17. Lou Y, Qesmi R, Wang Q, Steben M, Wu J, Heffernan JM. Epidemiological impact of a genital herpes type 2 vaccine for young females. *PLoS ONE*. 2012;7(10):e46027. https://doi.org/10.1371/journal.pone.0046027
18. Olsson A, Sandberg G, Dahlblom O. On Latin hypercube sampling for structural reliability analysis. *Struct Safety*. 2003;25(1):47-68.https://doi.org/10.1016/s0167-4730(02)00039-5
19. Larrañaga P, Lozano JA, eds.. *Estimation of Distribution Algorithms*: Springer; 2002. https://doi.org/10.1007/978‐1-4615-1539-5
20. Bosman PAN, Thierens D. Expanding from discrete to continuous estimation of distribution algorithms: the IDEA. *Parallel Problem Solving From Nature PPSN VI*: Springer; 2000:767-776. https://doi.org/10.1007/3-540-45356-3_75
21. Costa M, Minisci E. MOPED: a multi-objective Parzen-based estimation of distribution algorithm for continuous problems. *Evolutionary Multi-Criterion Optimization*, Lecture Notes in Computer Science: Springer; 2003:282-294. https://doi.org/10.1007/3-540-36970-8_20
22. Ding N, Zhou S, Sun Z. Optimizing continuous problems using estimation of distribution algorithm based on histogram model. *Simulated Evolution and Learning*, Lecture Notes in Computer Science: Springer; 2006:545-552. https://doi.org/10.1007/11903697_69
23. Marini F, Walczak B. Particle swarm optimization (PSO). A tutorial. *Chemom Intell Lab Syst*. 2015;149:153-165. https://doi.org/10.1016/j.chemolab.2015.08.020
24. Annual report of the microbiological information system 2012. N.I.P.O. 725–14–003–X. Sistema de Información Microbiológica. Centro Nacional de Epidemiología. Instituto de Salud Carlos III; Madrid, 2014. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/informesSIM/SIM2012.pdf
25. Annual report of the microbiological information system 2013. N.I.P.O. 725–15–0013. Sistema de Información Microbiológica. Centro Nacional de Epidemiología. Instituto de Salud Carlos III; Madrid, 2015. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/informesSIM/SIM2013.pdf
26. Annual report of the microbiological information system 2014. N.I.P.O. 725-16-0070. Sistema de Información Microbiológica. Centro Nacional de Epidemiología. Instituto de Salud Carlos III; Madrid, 2015. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/informesSIM/SIM2014.pdf
27. Annual report of the microbiological information system 2015. N.I.P.O. 062-17-0057. Sistema de Información Microbiológica. Centro Nacional de Epidemiología. Instituto de Salud Carlos III; Madrid, 2016. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/informesSIM/SIM2015.pdf

28. Annual report of the microbiological information system 2016. N.I.P.O. 062-17-0062. Sistema de Información Microbiológica. Centro Nacional de Epidemiología. Instituto de Salud Carlos III; Madrid, 2017. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/informesSIM/SIM2016.pdf
29. National Statistics Institute, Spain. https://www.ine.es/
30. Navarro Ortega D, Navalpotro Rodríguez D, Fraile Santos O. Actualización en el diagnóstico del herpes genital. Update on genital herpes diagnosis. https://www.seimc.org/contenidos/ccs/revisionestematicas/viromicromol/Herpesgen.pdf
31. Brauer F, Castillo-Chavez C. *Mathematical Models in Population Biology and Epidemiology*: Springer; 2012. https://doi.org/10.1007/978-1-4614-1686-9
32. Murray JD, ed. *Mathematical Biology*: Springer; 2002. https://doi.org/10.1007/b98868
33. Scott DW. *Multivariate Density Estimation: Theory, Practice, and Visualization*, 2nd ed. Wiley; 2014.
34. Numpy programming package. https://numpy.org
35. Scipy programming package. https://www.scipy.org/
36. Burgos C, Cortés JC, Martínez-Rodríguez D, Villanueva RJ. Computational modeling with uncertainty of frequent users of e-commerce in Spain using an age-group dynamic nonlinear model with varying size population. *Adv Complex Syst*. 2019;22(4):1950009. https://doi.org/10.1142/s0219525919500097
37. Makridakis S. Accuracy measures: theoretical and practical concerns. *Int J Forecast*. 1993;9(4):527-529. https://doi.org/10.1016/0169-2070(93)90079-3
38. Annual report of the microbiological information system 2017. N.I.P.O. 834-20-006-5. Sistema de Información Microbiológica. Centro Nacional de Epidemiología. Instituto de Salud Carlos III; Madrid, 2020. https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/informesSIM/Informe_Anual_SIM_2017.pdf

How to cite this article: Cortés J-C, Martínez-Rodríguez P, Moraño J-A, Romero J-V, Roselló M-D, Villanueva R-J. Probabilistic calibration and short-term prediction of the prevalence herpes simplex type 2: A transmission dynamics modelling approach. *Math Meth Appl Sci*. 2022;45(6):3345-3359. https://doi.org/10.1002/mma.7628