This is a post-peer-review, pre-copyedit version of a paper published in Cazzaniga P, Besozzi D, Merelli I & Manzoni L (eds.) *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Lecture Notes in Computer Science, 12313. CIBB 2019: 16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics, Bergamo, Italy, 04.09.2019-06.09.2019. Cham, Switzerland: Springer. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-63061-4_3

Effective use of evolutionary computation to parameterise an epidemiological model

Ryan Mitchell[1], David Cairns[1][0000−0002−0246−3821], Dalila Hamami[2][0000−0002−6616−3887], Kevin Pollock[3][0000−0003−0635−307X], and Carron Shankland[1][0000−0001−7672−2884]

[1] Computing Science and Mathematics, University of Stirling, carron.shankland@stir.ac.uk
[2] Université Abdelhamid Ibn Badis Mostaganem
[3] Programme Lead for MSc in Infection Prevention & Control, Inverness College, University of the Highlands and Islands

Abstract. Predictive epidemiological models are able to be used most effectively when they have first been shown to fit historical data. Finding the right parameters settings for a model is complex: the system is likely to be noisy, the data points may be sparse, and there may be many inter-related parameters. We apply computational intelligence and data mining techniques in novel ways to investigate this significant problem.
We construct an original computational model of human papilloma virus and cervical intraepithelial neoplasia with the ultimate aim of predicting the outcomes of varying control techniques (e.g. vaccination, screening, treatment, quarantine). Two computational intelligence techniques (genetic algorithms and particle swarm optimisation) are used over one-stage and two-stage optimisations for eight real-valued model parameters. Rigorous comparison over a variety of quantitative measures demonstrates the explorative nature of the genetic algorithm (useful in this parameter space to support the modeller). Correlations between parameters are drawn out that might otherwise be missed. Clustering highlights the uniformity of the best genetic algorithm results.
Prediction of gender-neutral vaccination with the tuned model suggests elimination of the virus across vaccinated and cross-protected strains, supporting recent Scottish government policy. This preliminary study lays the foundation for more widespread use of computational intelligence techniques in epidemiological modelling.

Keywords: genetic algorithm · particle swarm optimisation · epidemiology · human papilloma virus · k-means clustering.

1 Scientific Background

Computational models are increasingly used to investigate complex biological systems such as disease spread and the effects of interventions such as vaccination and treatment. Choosing parameters for these models can be challenging. A

small change to a parameter value or the model structure can be the difference between a large outbreak, endemic disease, or elimination of the disease. Computational intelligence approaches are highly suitable tools to support modellers to fit a model to historical data, as shown in our previous work [1], and provide a base for future predictions. We compare and contrast two such tools (genetic algorithm (GA) and particle swarm optimisation (PSO)) in a realistic case study of international importance. In addition, we extend our use of clustering techniques with process algebra [2] to investigate the quality of the results of the optimisation.

The Human Papillomavirus (HPV) is a sexually transmitted infection which affects the skin and moist membranes of humans. There are over 200 strains of this virus which are split into two subcategories known as oncogenic (high-risk) and non-oncogenic (low-risk). Oncogenic strains of this virus play a pivotal role in the development of various cell abnormalities including tumours [3]. Almost all cases of cervical cancer in women are attributed to high-risk HPV strains [4]. The World Health Organisation has made a global call to eliminate cervical cancer worldwide [5], driven through better understanding of HPV control measures.

This paper contributes to that understanding by presenting computational intelligence and data mining approaches to optimising a model of HPV and pre-cancerous stages. The model is fitted to Scottish data to demonstrate the techniques. According to Kavanagh et al. [6] 80% of cervical cancer cases in Scotland are attributed to HPV 16 & 18. Females aged 12-13 have received a bivalent vaccine against those strains since September 1st 2008 in Scotland. This programme has reduced HPV 16 & 18 prevalence from 30% to 4% in females [6]. The bivalent vaccine provides cross-protection against strains HPV 31, 33 & 45; however, that protection may reduce over time. Predictive modelling is used to investigate equal vaccination policies and the effect of waning cross-protective immunity.

As a starting point, we use a real-valued Genetic Algorithm (GA) as a general-purpose heuristic solver and Particle Swarm Optimisation (PSO) as an alternative solver that is useful for searching continuous parameter spaces. The model has eight parameters, which is a modest challenge for the chosen optimisation techniques but has produced some unexpected observations of the correlations between the parameters. The relationships between the optimal solutions within and between each approach are further explored through the use of data mining (clustering).

We use the optimised model to predict future outcomes of variants of the vaccination programme. We predict that a gender-neutral vaccination program can eradicate high-risk oncogenic strains of HPV even where immunity to those strains wanes. This work establishes a framework for applying computational intelligence to process algebra modelling, with effective and rigorous approaches to evaluating the resulting solution set.

2 Materials and Methods

2.1 Tools

The Evolving Process Algebra toolkit (EPA) [1] was developed with the goal of bridging the gap between evolutionary computation and the formal modelling domain, specifically using process algebra for modelling due to its parsimony and suitability for biological systems of interacting components. EPA is built on the ECJ Evolutionary Computation toolkit and the Bio-PEPA (deterministic or stochastic) simulation engine [7]. For this study, EPA was extended with a parameter sweep component and a particle swarm optimisation (PSO) module, complementing the existing GA approach to optimisation of numeric model parameters. EPA is able to optimise model structure and numeric parameters [1]. Here, the core aspects of the model are standard and only numeric parameters need to be optimised.

2.2 Model

The HPV model is a classic SEIR (Susceptible, Exposed, Infected, Recovered (Immune)) compartmental model [8] extended with binary gender and the potential precursor stages of cervical cancer (three stages of cervical intraepithelial neoplasia (CIN). Epidemiological studies [6, 9] informed model development. Fig. 1 shows the compartments of the model and the available transitions between compartments. The Bio-PEPA code for the model is archived online [10].

There are over 200 strains of HPV. For this investigation our model groups together the two oncogenic strains HPV16 and HPV18 which are targeted by the bivalent vaccine. The cross-protected strains HPV 31, 33 and 45 (some level of protection is conferred by the bivalent vaccine) are identified as "other" in the model. Further strains of HPV have been omitted.

Vaccination has been added in a staged manner via a rolling vaccination programme for girls aged 12-13 and a targetted catch-up programme for older girls (14-18) (who may have already been exposed to HPV strains, so vaccination may be unsuccessful). The catch-up programme ran for 3 years; in 2014, the eligible routine age for vaccinations was amended to 11-12 year olds for logistical purposes. The model gradually increases both vaccine takeup and vaccine efficacy to capture these features. The timescale is one simulation tick per day, therefore rate constants are expressed as daily rates. The population is open, with births, deaths and immigration. The birth rate is chosen to balance up the death rate in the Scottish population (keeping the population constant for the duration of the study). Immigration (average number of imported infections per year) is given by the standard formula $0.02\sqrt{N}$, where N is the total population [13]. For optimisation, the simulation time is seven years, matching the data of Kavanagh et al [6]. The vaccine is introduced after the first simulation year.

HPV is a sexually-transmitted infection. Couplings are simplified as being female-female, female-male, male-female, and male-male, in proportions reflecting sexual orientation demographics in Scotland [11]. Coupling preference is not

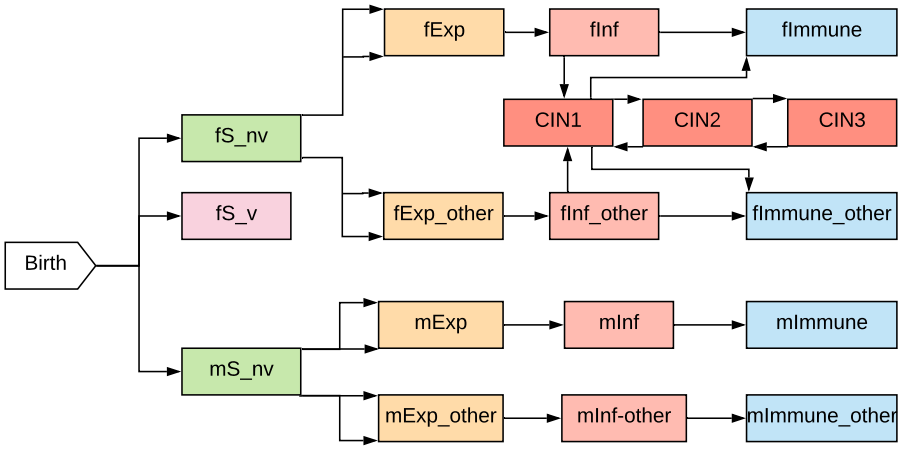


Fig. 1. Schematic of the HPV model showing compartments and routes between compartments. Birth is a source of new individuals. The compartments S, Exp, Inf, Immune are prefixed f (females) and m (males). The suffix _v or _nv relates to vaccinated or not vaccinated. HPV 16 and 18 are the Exp → Inf → Immune strands, and HPV 31, 33 and 45 have the suffix other. Three pre-cancerous stages CIN1, CIN2, CIN3 affect females only.

fixed, reflecting fluidity of orientation and behaviours. This simplification is a pragmatic modelling choice suiting the Markov chain semantics of Bio-PEPA. The two routes from susceptible to exposed in Fig. 1, e.g. fS_nv to fExp reflect that the infection can come from a female or a male partner.

HPV can be described as using frequency-dependent transmission: the number of infectious contacts an individual can make in a single day is proportional to SI/N, where S and I are the number of susceptible and infected individuals respectively and N is the total population. We assume an average of 3.5 days between exposure and becoming infectious [12].

2.3 Optimisation

The rates of progression and regression between the CIN stages is unknown, therefore these were targeted for optimisation together with the contact rate (Cr) and the rate of developing immunity to strains 16 & 18 (Immunity). Domain knowledge was used to set appropriate upper and lower bounds for these rates. The target data was extracted from Kavanagh et al. [6] (HPV, 7 years population sampling) and Pollock et al [9] (CIN, 5 years population sampling).

Four separate experiments were carried out:

1. two-stage optimisation where infection parameters Cr, Immunity were obtained by a parameter sweep followed by CIN parameters CIN1, CIN1r, CIN2, CIN2r CIN3, CIN3r obtained using a GA,

2. two-stage optimisation as 1. but using PSO,
3. one-stage optimisation of all eight unknown parameters using a GA,
4. one-stage optimisation as 3. using PSO.

Our goal is to explore the potential use of GA and PSO with algorithm parameters that allow for a balance of exploration versus exploitation and that could be completed on our local compute cluster in a reasonable time (e.g. 1 day). The GA used a population of 100 individuals over 100 generations, with tournament selection (5% tournament pool), 1% elitism, 1-point crossover (100%), 5% mutation rate, and generational replacement. The PSO used 100 epochs, 100 particles, 0.728844 inertia, 1.49 cognitive component and 1.49 social component. For this comparison standard parameter settings were used. Ideally we would adjust the GA/PSO parameters to make the search process more efficient. We have moderated for this by using long runs and repeating the runs. Given the number of independent valid solutions we generated that converged to similar fitness scores, we can be reasonably confident that our solution set is appropriate. Optimising the algorithm parameters first may have resulted in quicker convergence but this had to be balanced with the time taken to find these best-case algorithm parameters. If our solution set had been more diverse and had not been a good fit to the data then we would have refined our algorithm parameters.

For both the GA and the PSO, inputs to be optimised were internally represented by real values as used by the HPV model. Prevalence data for HPV in Scotland [6] provides relevant target values at fixed points over time. For both optimisers, fitness was calculated as the Euclidean distance score between the data and the equivalent predicted model values. This was chosen as a straightforward standard linear measure for this data. Fixed data points in years 2009, 2010, 2012, 2013 and 2015 were used. See Figs. 2 and 3 for the fixed point values and a comparison trace produced by a high fitness simulation. For each algorithm, results were collated over 100 independent runs (i.e. 10,000 solutions), producing 100 top solutions. From this collated subset of 100 solutions, the top 20 were analysed.

3 Results

3.1 Parameter Optimisation

To illustrate optimisation results against historical data, the best match trace is shown in Figs. 2 and 3 (GA two-stage results). Mean, median and standard deviation for all eight parameters across twenty best-fit solutions are shown in Tab. 1, comparing GA and PSO results, and two- and one-stage processes.

Results are largely equivalent in terms of fit but there is considerable variation in some parameter values in these solutions. Distribution and correlation information (Figs. 4–7) draw out the model's susceptibility to variance in these parameters. Where there is a strong correlation, the model produces a consistent response and is sensitive to a parameter's value. Where there is little to no correlation, the model is insensitive to the parameter with respect to overall fitness.

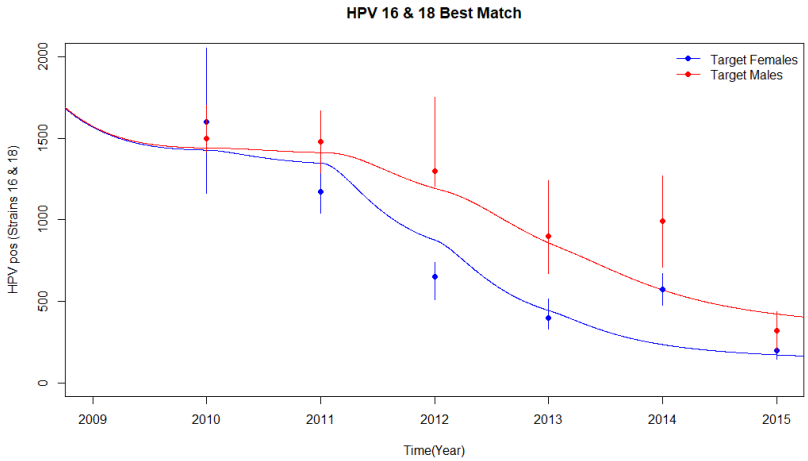


Fig. 2. A single, best-match stochastic simulation of HPV 16 & 18 infections parameterised from the two-stage process showing the data points and 95% confidence intervals of the original data

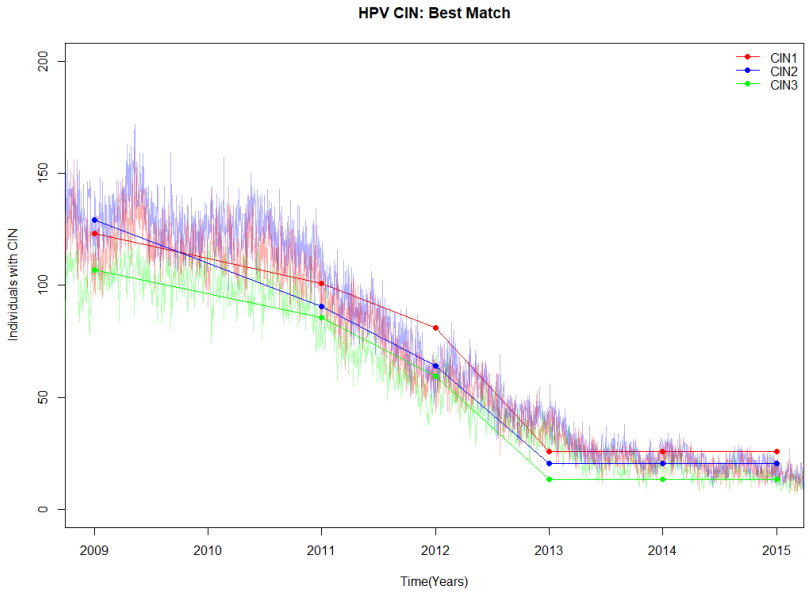


Fig. 3. A single, best-match stochastic simulation of CIN stages parameterised from the two-stage process

Table 1. Optimised parameter values with mean and standard deviation, comparing two- and one-stage, and GA and PSO (see also Figs. 4–7 diagonals).

	Bounds		Two Stage			
Parameter	Lower	Upper	GA		PSO	
Cr	0.03	0.06	0.0420	±0.00045	0.0420	±0.00045
Immunity	0.00014	0.0204	0.0028	±0.00003	0.0028	±0.00003
CIN1	0.01	0.90	0.65	±0.206	0.88	±0.019
CIN2	0.01	0.90	0.02	±0.013	0.54	±0.192
CIN3	0.01	0.90	0.45	±0.288	0.79	±0.036
CIN1r	0.01	0.90	0.39	±0.263	0.47	±0.286
CIN2r	0.01	0.90	0.39	±0.212	0.79	±0.037
CIN3r	0.01	0.90	0.46	±0.254	0.50	±0.286

	Bounds		One Stage			
Parameter	Lower	Upper	GA		PSO	
Cr	0.03	0.06	0.0500	±0.00005	0.049	±0.00065
Immunity	0.00014	0.0204	0.0015	±0.00045	0.003	±0.00004
CIN1	0.01	0.90	0.90	±0.127	0.90	±0.001
CIN2	0.01	0.90	0.01	±0.007	0.34	±0.322
CIN3	0.01	0.90	0.51	±0.238	0.90	±0.002
CIN1r	0.01	0.90	0.46	±0.224	0.59	±0.323
CIN2r	0.01	0.90	0.49	±0.223	0.80	±0.227
CIN3r	0.01	0.90	0.58	±0.268	0.53	±0.265

Figs. 4–7 illustrate the relationship between fitness and the other variables for GA: two-stage, GA: one-stage, PSO: two-stage and PSO: one-stage respectively. These figures are quite complex, so we explain the format here, and repeat the main features in the captions to aid the reader.

The x-axes labels alternate on the top and the bottom, while the y-axes labels alternate on the left and the right. The variables under optimisation are labelled along the diagonal from top left to bottom right. Each position on this diagonal includes a histogram of the named variable's distribution of values in the solution set of top twenty results. For example, the distribution of CIN1r is skewed to the left in Fig. 4 but to the right in Figs. 5–7.

The bottom left section of the figures show scatter plots of pairs of variables. Track vertically and horizontally to locate the relevant pair: for example, in Fig. 4 the scatter plot second row from top, first column from left shows strong correlation of fitness against CIN1.

The upper right side of the figure shows the corresponding correlation values for the scatter plots from the lower left section. For example, in Fig. 4 the plot of CIN2 against CIN2r (5th row, 4th column) shows a strong correlation, and this matches a correlation value of 1.00 (4th row, 5th column).

Figs. 4–7 highlight robustness, i.e. likely model response to small changes in parameter values. While both the GA and PSO approaches have some variables with low pairwise correlations, the GA had slightly better variable to fitness correlation values. Results were broadly similar for the GA across the 1-stage and 2-stage tasks: adding extra variables did not significantly vary the results.

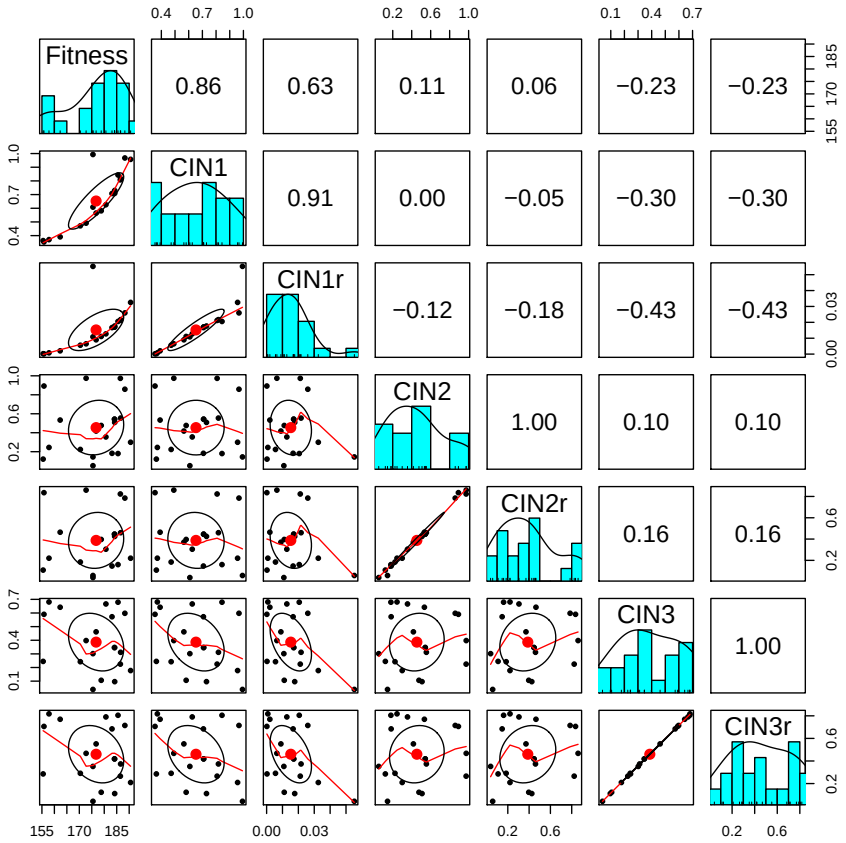


Fig. 4. GA parameter correlations: two-stage. Variables under optimisation are labelled along the diagonal from top left to bottom right. Each position on this diagonal includes a histogram of the named variable's distribution of values. The bottom left section of the figure shows scatter plots of pairs of variables. The upper right side of the figure shows the corresponding correlation values for those scatter plots. The x-axes labels alternate on the top and the bottom, while the y-axes labels alternate on the left and the right.

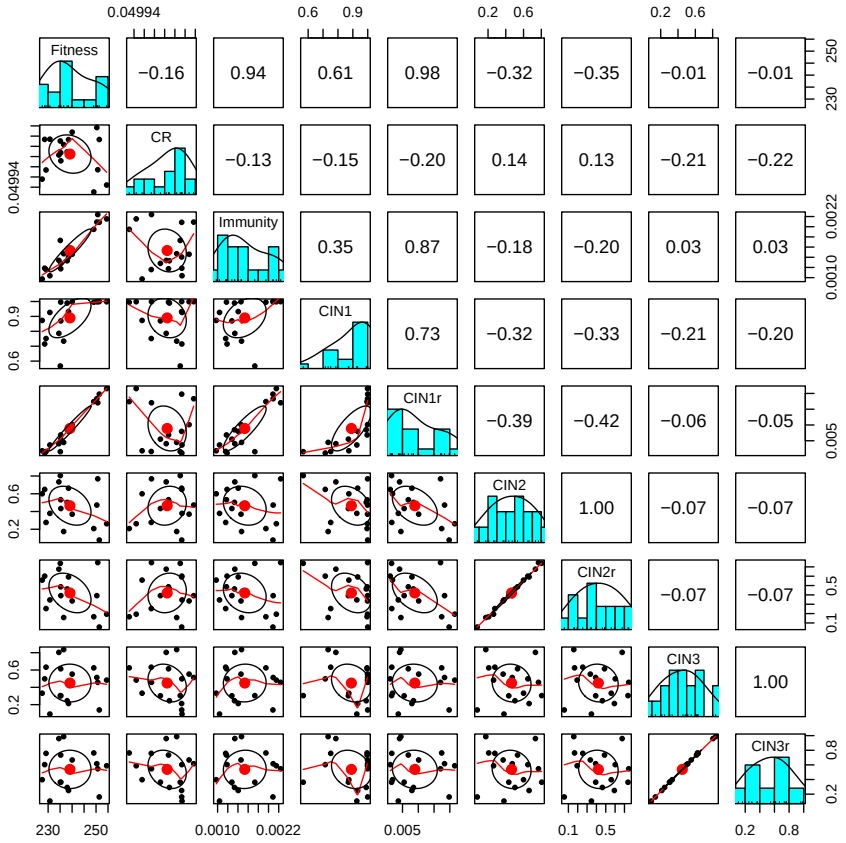


Fig. 5. GA parameter correlations: one-stage. Variables under optimisation along the diagonal. Bottom left shows scatter plots of pairs of variables. Upper right side shows the corresponding correlation values. X-axes labels alternate top and bottom. Y-axes labels alternate on left and right.

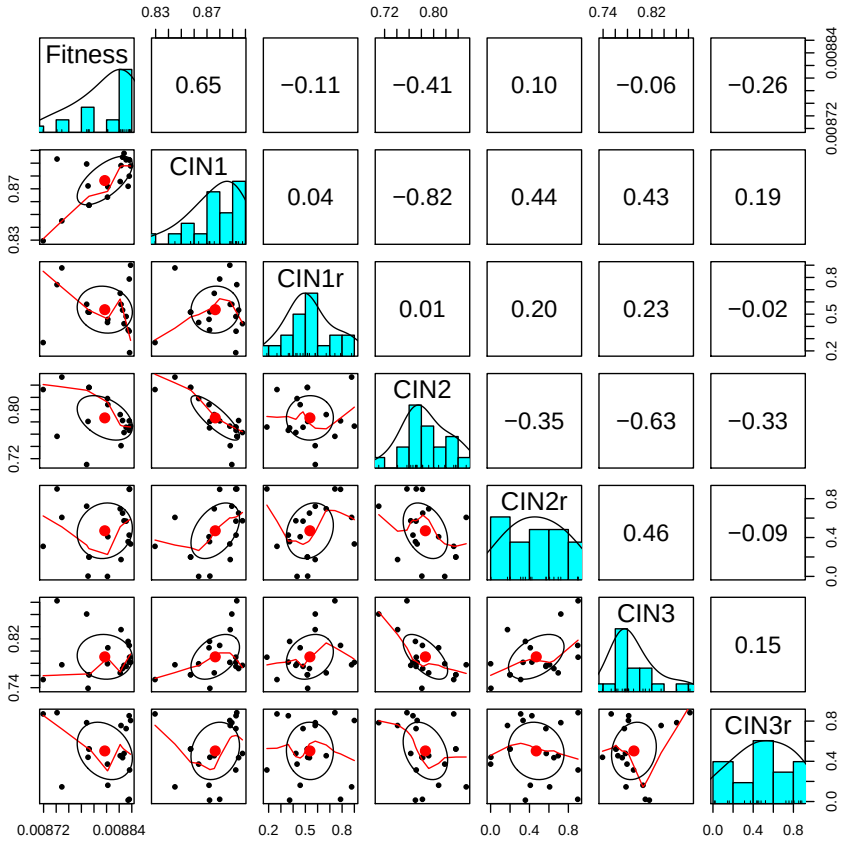


Fig. 6. PSO parameter correlations: two-stage. Variables under optimisation along the diagonal. Bottom left shows scatter plots of pairs of variables. Upper right side shows the corresponding correlation values. X-axes labels alternate top and bottom. Y-axes labels alternate on left and right.

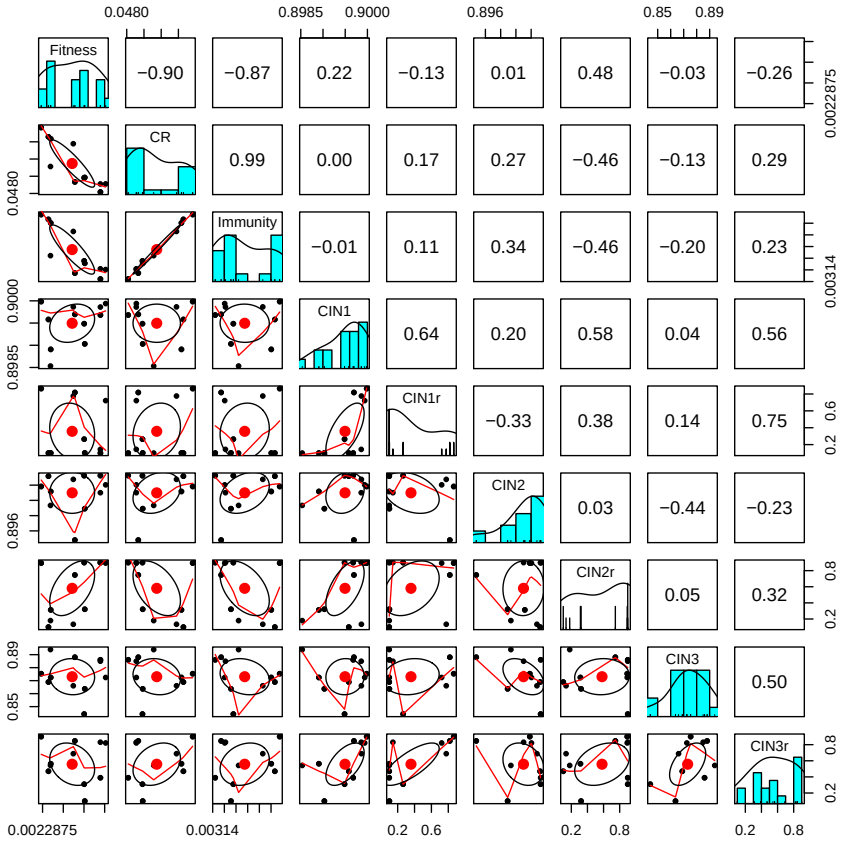


Fig. 7. PSO parameter correlations: one-stage. Variables under optimisation along the diagonal. Bottom left shows scatter plots of pairs of variables. Upper right side shows the corresponding correlation values. X-axes labels alternate top and bottom. Y-axes labels alternate on left and right.

In contrast, the PSO seems more sensitive to the Cr and Immunity parameters in the 2-stage process.

3.2 Clustering the Optimal Solutions

Several experiments were carried out using k-means clustering to further investigate the relationships between the produced optimal solutions. Clustering was applied to analyse the stability of the GA results only, and of the PSO results only. Finally, clustering was applied to GA and PSO results across the 1-stage or 2-stage process to demonstrate which technique produced more stable solutions. Values of k from 2-5 were tried. Euclidean distance was used as the distance function, with k=3 producing optimal clusters. Distribution of data points across clusters were monitored to provide insight into how k-means splits the data for each k and for each optimisation approach. Cluster sizes are more unequal for k=3. For the other values 2,4 and 5 the number of instances are more equally distributed across clusters and therefore do not provide any additional information.

Tab. 2 shows the results of using k-means clustering (k=3) on the combined GA and PSO results for 1-stage or 2-stage optimisation. Clustering across the approaches highlights the similarity of solutions produced within each method. All 20 GA results are in cluster 0 (1-stage optimisation). All 20 PSO instances are in cluster 0 (2-stage optimisation), highlighting sensitivity to Cr, Immunity.

Table 2. Clustering of GA and PSO solutions (1-stage left, 2-stage right).

	Centroids							
	1-stage GA&PSO solutions				2-stage GA&PSO solutions			
Attribute	All	Cluster0	Cluster1	Cluster2	All	Cluster0	Cluster1	Cluster2
Cr	0.0494	0.05	0.0484	0.0496	0.042	0.042	0.042	0.042
Immunity	0.0023	0.0015	0.0032	0.0032	0.0028	0.0028	0.0028	0.0028
CIN1	0.90	0.90	0.90	0.90	0.76	0.56	0.81	0.88
CIN1r	0.17	0.0083	0.31	0.38	0.27	0.0092	0.0266	0.51
CIN2	0.71	0.51	0.90	0.90	0.62	0.41	0.49	0.79
CIN2r	0.53	0.46	0.75	0.37	0.43	0.35	0.40	0.49
CIN3	0.64	0.49	0.75	0.87	0.58	0.47	0.16	0.78
CIN3r	0.56	0.58	0.46	0.64	0.48	0.57	0.19	0.51

3.3 Model Predictions

Having optimised the model to match data [6, 9], predictions of vaccination policies can be made. For example, a gender-neutral policy of vaccination is clearly equitable. Fig. 8 shows the predicted results of the gender-neutral policy and how it produces elimination of the vaccine-specific virus strains, elimination of cross-protected strains, and subsequent reduction of CIN with concomitant decrease

in cervical cancer. As gender-neutral vaccination was partly (due to COVID-19) implemented in Scotland from 2019 these predictions will soon be able to be tested against observed data.

With no vaccination, the model suggests half of the population will be infected with HPV strains 16 & 18, and 31, 33 & 45. Female-only vaccination, shown in Fig. 8 (a), reduces HPV 16 & 18 prevalence to 6% (female) and 15% (male) of the population. In contrast, Fig. 8 (c) shows these strains are virtually eliminated with equal vaccination. Fig. 8 (a, c) shows that strains HPV 31, 33 & 45 also reduce due to cross-protection from the bivalent vaccine. If that cross-protection wanes over time, e.g. Fig. 8 (b) shows 9.4 years protection, then strains 31, 33 & 45 become dominant in a female-only vaccination regime with 10% females and 23% males infected. Equal vaccination eliminates these strains, shown in Fig. 8 (d).

4 Conclusion

Two computational intelligence techniques were used to refine a hand-built model of HPV and subsequent CIN stages. We originally took a two-stage approach with two different optimisation algorithms to cross-compare results and parameter variance. A further one-stage 8-parameter optimisation run for the the GA and PSO approaches was performed at the end of the study to investigate if this would lead to similar results. Clustering was used to provide further insight into the results generated.

Comparing GA performance to PSO performance, we see, as expected, the GA is more explorative (slightly wider range of fit results in Tab. 1 and Figs. 4 and 5) and a good general-purpose heuristic. The one-stage process provides strong correlations between CIN progression and regression parameters. The GA pulled out an interesting and unexpected strong positive correlation between Immunity and CIN1r: the state fImmune can be reached either directly from fInf, or via CIN1. Clustering shows the similarity of top GA results for the one-stage process.

For PSO the solutions (Figs. 6 and 7) showed no clear correlations between parameters except Cr and Immunity in the one-stage optimisation. Variance was lower overall in PSO results between one-stage and two-stage experiments. Clustering shows that the top PSO results are similar for the two-stage process: PSO gives a more uniform solution set when additional constraints are applied.

In future work, a range of algorithm parameters could be explored to improve GA or PSO performance. GA performance could most obviously be improved by running for longer, and decreasing mutation as the number of generations increases. The current work sets a baseline for such experiments.

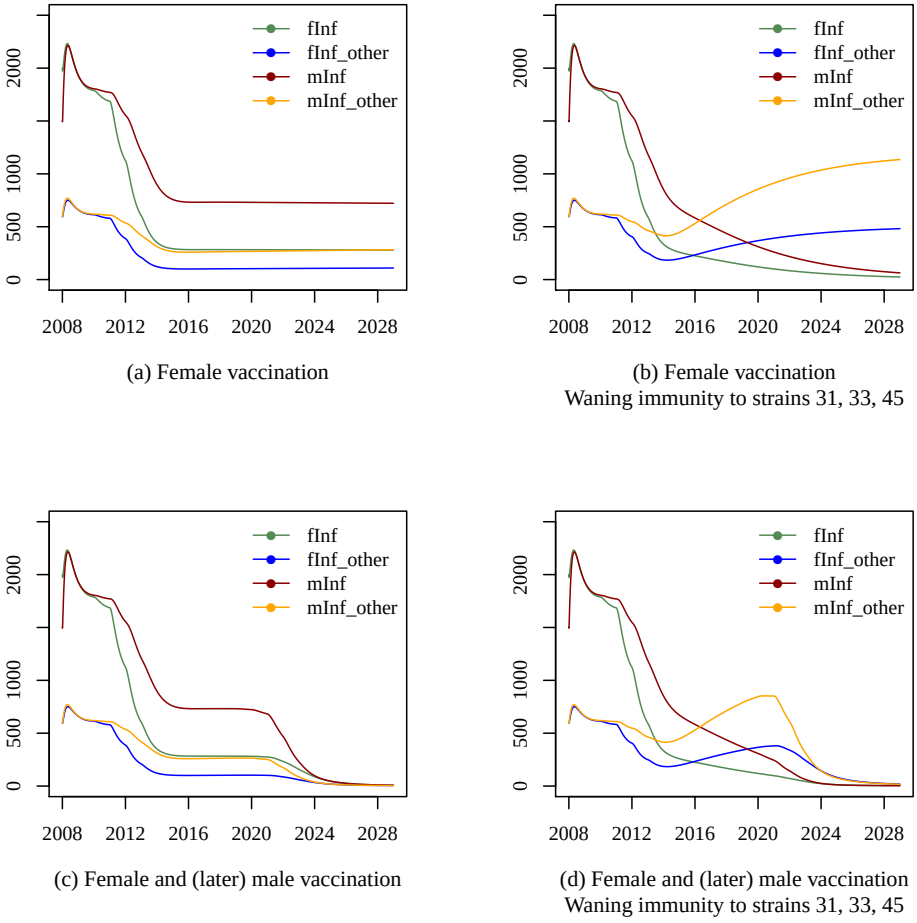
The refined model can now be used to make computational predictions from health care data sets relating to gender policies on the bivalent vaccine. Varying vaccine efficacy or uptake would determine potential critical vaccination thresholds to ensure long-term eradication of relevant cancers. To calibrate our model we have used data relating to HPV in Scotland; however, due to a worldwide

effort to eliminate cervical cancer, several data sets are available to validate such models. Future work will develop our model to apply to a range of data sets, testing predictions across vaccination scenarios, allowing more robust statistical analysis of the results. This would provide a contrast with the results of Simms et al [14].

Much further predictive work is possible if the model is extended, for example, with additional HPV strains and different vaccines. In addition, there is evidence that HPV vaccination for males can reduce head and neck cancers. Given appropriate data, it would be of value to show the impact of vaccination and screening on both males and females with respect to a variety of cancers.

References

1. D. Marco, D. Cairns and C. Shankland, "Optimisation of process algebra models using evolutionary computation" *in 2011 IEEE Congress of Evolutionary Computation (CEC)*, pp. 1296–1301, 2011.
2. D. Hamami et al., "Improving process algebra model structure and parameters in infectious disease epidemiology through data mining". *J. Intelligent Information Systems*: 52 (3), pp. 477–499, 2019. https://doi.org/s10844-017-0476-1
3. R. Faridi et al, "Oncogenic potential of Human Papillomavirus (HPV) and its relation with cervical cancer." *Virology Journal*, vol. 8, pp. 269, BioMed Central.
4. FX. Bosch, A. Lorincz, N. Muñoz, CJLM. Meijer and KV. Shah. "The causal relation between human papillomavirus and cervical cancer" *in J Clin Pathol.* 2002;55:244–265. https://doi.org/10.1136/jcp.55.4.244
5. M. Brisson and M. Drolet, "Global elimination of cervical cancer as a public health problem". *Lancet oncology* 20(3):319–321; 2019.
6. K. Kavanagh, K.G. Pollock, et al., "Changes in the prevalence of human papillomavirus following a national bivalent human papillomavirus vaccination programme in Scotland: a 7-year cross-sectional study". *The Lancet Infectious Diseases*, 17 (12):1293–1302, 2017.
7. F. Ciocchetta, A. Duguid, S. Gilmore, M.L. Guerriero, and J. Hillston. The Bio-PEPA Tool Suite. http://homepages.inf.ed.ac.uk/jeh/Bio-PEPA/, 2009.
8. R. Anderson and R. May, "Population biology of infectious-diseases". *Nature* 280: 361–367; 1979.
9. K.G. Pollock, K. Kavanagh, et al., "Reduction of low- and high-grade cervical cancer abnormalities associated with high uptake of the HPV bivalent vaccine in Scotland". *BJC* 1-7, 2014.
10. System Dynamics model archive, www.cs.stir.ac.uk/SystemDynamics/models/
11. Sexual Orientation Demographics. Scottish Government, Dec 2018.
12. J.T. Schiller et al, "Current understanding of the mechanism of HPV infection". *Gynecologic oncology* vol. 118,1 Suppl: S12-7, 2010.
13. B.F. Finkenstädt, M. Keeling, B.T. Grenfell, "Patterns of density dependence in measles dynamics". *Proceedings of the Royal Society B* 265: 753–762, 1998.
14. K. Simms et al, "Impact of scaled up human papillomavirus vaccination and cervical screening and the potential for global elimination of cervical cancer in 181 countries, 2020-99: a modelling study". *Lancet Oncology*: 20(3) pp. 394–407 2019. https://doi.org/10.1016/S1470-2045(18)30836-2.



(a) Female vaccination

(b) Female vaccination
Waning immunity to strains 31, 33, 45

(c) Female and (later) male vaccination

(d) Female and (later) male vaccination
Waning immunity to strains 31, 33, 45

Fig. 8. Predicted impact on HPV 16 & 18, and cross-protected strains HPV 31, 33 & 45 generated by Ordinary Differential Equation-based simulation. Female vaccination from 2008 (all plots). Equal vaccination from 2019 (c, d). Waning immunity to cross-protected strains (b, d)