19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Application of Microsimulation to Disease Transmission and Control

A. Green[a], **Y. Zhang**[a], **I. Piper**[a] **and D. Keep**[a]

[a] *University of Wollongong, Northfields Ave, Wollongong, NSW 2500*

Email: tgreen@uow.edu.au

Abstract: In this paper, we present an extension of our novel microsimulation technique, as applied to biological infection spread, to estimate the underlying causal parameters driving an infectious process. The underlying simulation framework, Simulacron, was developed in order to understand the development and course of, response to and recovery from single and multiple threats on community populations. Such threats include a range of natural (such as disease spread in communities, fire, flood etc.) and manmade events (such as terrorism, including the use of biological agents, money laundering, smuggling as well as accidents etc.). These threats can cause serious disruption to modern society and the optimal approaches to prevention, mitigation, response and recovery are little understood. Furthermore, assumptions currently used cannot otherwise be readily tested.

A case study of the 1920 influenza outbreak at the Royal Naval School, Greenwich has previously been used to demonstrate the feasibility of the simulator. This case is well documented and has detailed information about the typical education schedules and the physical locations within the school as well as documentation on the disease spread. It also has the advantage of being a semi-closed environment with about 1,000 pupils at any one time in the school; a manageable size for simulation on a single processor. One significant issue which arose during this study is the difficulty of equating traditional epidemiological measures of disease virulence (reproduction number, etc.) with the causal parameters (infection timings, cross-infection probability, etc.) used in our model. This paper describes two programs, developed for use with the simulator, to estimate the causal parameters that best fit these traditional measures. *Refinery*, given an initial range for each causal parameter, performs Monte Carlo sampling to produce a set of candidate parameter instances. Refinery then performs simulated annealing in order to refine the causal parameter estimate ranges. *Monotony* then performs multiple simulations for each of these instances, varying only in random seed. Once complete, values for the statistical measures are computed for each instance.

Three results from the simulation in which the infection parameters and the number of Susceptibles in the population were chosen randomly from a normal population are shown in Figure 1. The 1920 outbreak is superimposed assuming one infection cycle as an offset. It is possible that one case went undetected prior to the historical observations of the outbreak in those runs shown. The mean time to isolation from infection was 41.7±0.2 hours and the mean time to infection by a person already infected was 30.2±2.9 hours.

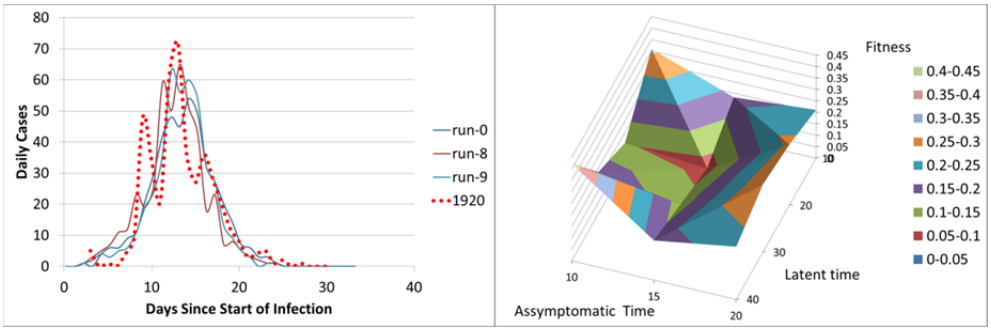

Figure 1: The closest fits to the historical data: a) daily cases b) local minimum of "fitness"

The reproduction number calculated by this method agrees with a simple SIR calculation on the historical data. The simulation gives an average hero time of 7 hours before discovery and isolation. As a result the infectious time is 22 hours which is similar to the 24 hours in the literature. The one parameter that does not match assumptions on influenza data is the latent time. The simulation is giving a value of 20 hours compared to literature values of 35 to 45 hours. Figure 1b shows the minimum fitness as a function of latent and asymptomatic times indicating that the results obtained sit in a minimum that practically rules out other values for the latent time.

Keywords: *Epidemiology, Microsimulation, Decision Support Systems*

1 INTRODUCTION

In this paper we discuss the application of microsimulation to epidemiology and how to derive parameters that are useful for describing infection spread. Since they were first introduced [Kermack et al. 1927] traditional epidemiological models have relied on parameterised rate equations for describing the spread of disease such as influenza. These rate equations have been gradually compartmentalised, either by age, or by social groups such as communities and families [Ferguson et al., 2005] leading to increasing demands for data that can be interpreted for the larger numbers of parameters required to solve the equations. Although simulation at the level of the individual has been used for some epidemiological studies [Connor et al., 2000], there is a perception that simulation describing individuals who mix socially moving through spatial locations is too time consuming to be worthwhile because of the inherent stochastic nature of these simulations that require multiple simulations to obtain average values.

Microsimulation potentially has a number of advantages over existing techniques in studying infection spread due to the inherent uncoupling of the contact time from the chance of infection. Some of the effects that can be modelled include:

- The chance that infection spread will occur with a particular disease through a population.
- The attack rate that would be expected when infection does spread.
- The different characteristics of Susceptibles that spread the disease from those who don't.
- How the disease spread is modified by individual behaviours and contact times.
- How spread through many hosts, before impacting on a specific target host (human or other species), can affect the impact.
- How the behaviour of individuals and other human factors impact infection spread. For example: delivery of vaccines to a hospital or the degree of hygiene that might be involved in spread.
- The risks involved in policy for infection control and how it compares with other types of risks in society.
- The ability to study infection spread based on properties of virus shedding, immune response and DNA tags.

One significant issue which arose during the development of an infection model is the difficulty of equating traditional epidemiological measures of disease (reproduction number, attack rate, virulence etc.) with the causal parameters (infection time constants, cross-infection probability, etc.) used in our model. The focus of this paper is to demonstrate a monte-carlo approach to equate traditional methods of epidemiology with results from microsimulation.

1.1 Simulacron

Simulacron is the microsimulation framework that we have developed. It has been designed to be scalable, to maintain a real time response, and to allow flexibility in the types of problem it can be used for studying. The simulation is based on two concepts: "*cells*", an abstract location and "*peeps*", a simulated individual who can be a person, a speck of dust, a particle, or animal—anything with a location.

Both *cells* and *peeps* can have arbitrary data fields associated with them to allow them to interact. These include names, infection states, schedules and behaviour patterns associated with other states. These fields are built from a series of templates that are designed to mimic the real world for the problem in question and which are instantiated to the whole population being simulated. This allows construction of behaviours that go from completely random to completely deterministic according to the problem being studied. In addition, the simulation requires modules which provide functionality such as moving peeps according to their individual schedule and changing the infection status of individuals. Extensive reports can be produced for subsequent analysis.

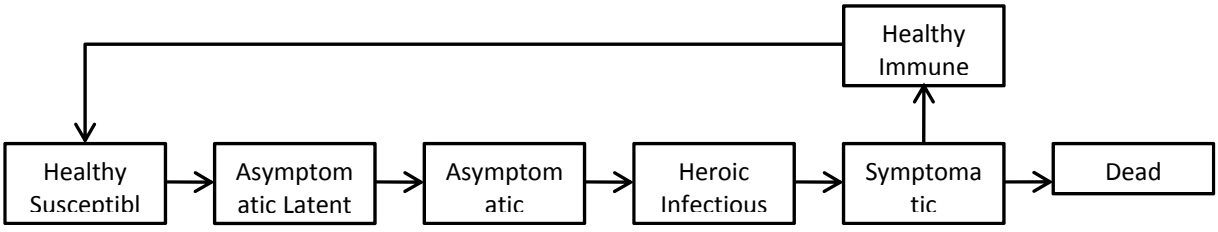

Figure 2: The infection process model used in Simulacron

The infection model used is shown in Figure 2. The infection spread occurs by proximity within the same cell with a probability of infection shown in Equation 1. Once infected, progress of an individual though each successive state is dictated by a time constant picked from a normal distribution through the population. The majority of epidemiological models assume a constant rate of recovery no matter the time from infection and hence rely on gamma distributions to describe the time periods for infection. As our method explicitly models contact between individuals there is no need for this assumption or mathematical convenience. Furthermore, recent studies have questioned the suitability of gamma functions in stochastic studies [Vergu et al, 2009]. Death is also a probability. Heroic time emulates a person who is symptomatic but continues with their normal schedule rather than self isolates or changes behaviour such as visiting the doctor. While more than one vector of disease spread can be studied by the use of a mask for the population and for cross-infection, multiple diseases within the same person cannot currently be modelled.

At each time step, an infection probability p_i, calculated for each peep i, is compared with a uniform [0,1] random number, r_i. If $r_i < p_i$ the peep becomes infected. The value of p_i is given in equation (1) where c_i is the chance of infection of the peep for the disease (relative to τ_i) which is a function of the disease and individual. N_i is the number of infected peeps in the cell, τ_i is the average infection time for the disease (the global temporal scaling for c_i) and Δt is the duration of the time step.

$$p_i = \frac{c_i N_i \Delta t}{\tau_i} \tag{1}$$

The simulation data sets are created by defining a series of templates for the description of the community and the behaviours that are desired. The community requires the specification of workplaces, households, attributes of the population together with daily and weekly schedules that align with the characteristics of the population. Data for this process generally comes from population census data and other studies. The behaviours can be personal behaviours such as self-isolation, hospital and doctor visits on becoming symptomatic or increased use of hygiene practices such as hand washing and personal social distancing strategies. They can also be social or desired policy behaviours such as school closures, forced isolation, prophylaxis and vaccination. Part of the simulation can also include logistic problems such as delivery of vaccination to the population, the desirability of setting up of fever hospitals as a pragmatic solution in pandemics and many more.

Two programs have been written: *Refinery*, which converges on the best solution to either historical or surveillance data through multiple instantiations of the simulation and *Monotony*, which provides the statistical mean and variance for a given instantiation.

1.2 Refinery and Monotony

P(t) represents a cumulative logistic function and is the closest theoretical curve for a population subject to depletion (in this case becoming immune as infection takes hold) [Modis, 1992]. The equation is parameterised to allow analysis of the data with the intent to match the moment, height and average slope of the cumulative curve. This curve fitting is relatively independent of fluctuations around the mean curve caused by waves of infection. The waves in our formulation are a direct consequence of the probabilistic nature of our model.

$$P(t) = \frac{K}{1+e^{-r(\tau+t)}} \tag{2}$$

$$\frac{d}{dt}P(t) = \frac{r}{K}P(t)[K - P(t)] \tag{3}$$

Where r is the inverse of the reproduction number, $R=zR_o$, z is the proportion of Susceptibles in the population and K is the number of people in the population who become infected. τ is the moment of the cumulative distribution curve, *P(t)*.

Equation 3 can be rearranged.

$$\frac{d}{dt}P(t) = rP(t) - \frac{r}{K}P^2(t) \tag{4}$$

Consider that, although P(t) is conceptually a continuous function, we are actually dealing with discrete values of t. Therefore, let t be taken from the sequence $t_1; t_2; \ldots; t_m$.

This allows us to rewrite the derivative as:

$$y = \beta X \tag{5}$$

where

$$y = \begin{vmatrix} \frac{d}{dt}P(t_1) \\ \frac{d}{dt}P(t_1) \\ \vdots \\ \frac{d}{dt}P(t_m) \end{vmatrix}, \qquad X = \begin{vmatrix} P(t_1) & P^2(t_1) \\ P(t_2) & P^2(t_2) \\ \vdots & \vdots \\ P(t_m) & P^2(t_m) \end{vmatrix} \quad \text{and} \quad \beta = \begin{vmatrix} r \\ -\frac{r}{K} \end{vmatrix} \tag{6}$$

A best estimate for β, $\hat{\beta}$, can be obtained using the following:

$$\hat{\beta} = (X^T X)^{-1} y \tag{7}$$

This yields, $\hat{r} = \hat{\beta}_{1,1}$ and $\hat{K} = \frac{-\hat{r}}{\hat{B}_{2,1}}$

The value of τ is found directly from the time at which the number of symptomatic individuals exceeds 50% of the total symptomatic individuals in *P(t)*.

Refinery implements this curve fitting at the end of each run. Runs are composed into iterations, where *n* individual runs are compared to the historical data to determine a fitness, f_n, for the iteration as a whole:

$$\Delta r_i = \frac{(r_{his} - r_i)}{r_{his}}; \Delta K_i = \frac{(K_{his} - K_i)}{K_{his}}; \Delta \tau_i = \frac{(\tau_{his} - \tau_i)}{\tau_{his}}; i = 1, n \tag{8}$$

$$f_i = \sqrt{\left(\Delta r_i^{\,2} + \Delta K_i^{\,2} + \Delta \tau_i^{\,2}\right)} \tag{9}$$

$$f_n = \min(f_i)\,; i = 1, n \tag{10}$$

New mean and standard deviations for the infection parameters are then computed from the run that most closely matches the data. A choice can be made as to whether or not to include non-infected people in value computations rather than just those who became infected. The refined values for the infection parameters are then used in a new iteration.

Once an estimate has been found, *Monotony* is then used to automatically produce the mean and standard deviations of an instance of the refined parameters. This is done by altering the random seed, but not an individual's infection parameters, in successive runs. Two sets of statistics are computed: the average of all runs and the average of those where infection spread occurs. The boundary between the two can be set as it was found that even when the majority of simulations showed large numbers of infection spread, there were a few in which the index case either did not infect anyone or only a few people.

2 THE 1920 INFLUENZA OUTBREAK AT THE ROYAL NAVAL SCHOOL, GREENWICH UK

This case study was chosen as it has well documented information about the typical education schedules and the physical locations within the school as well as documentation on the infection spread within the school [Bold, 2000; Dudley, 1926; Grist, 1976]. It also has the advantage of being a semi-closed environment with about 1000 pupils at any one time in the school — a manageable size for simulation on a single processor. The simplest model to infection was applied (Figure 2) rather than a more complex state model such as Mathews et al. [2007], to test the capability of this type of simulation.

The school housed nine dormitories that were a mix of all age groups. Classes during the day, however, were on the basis of age. Peeps were characterised by their dormitory and age that were used to develop a daily schedule that matched historical information. The infection parameters used at the start of all simulations are shown in column 2 of Table 1. The historical moment of the distribution curve, $\tau_{his,}$ is 9.5 days. The value of τ used in the simulation is τ_{his+} $\tau_{incubation}$, where $\tau_{incubation}$ is the time to isolation given initially by the latent, asymptomatic and hero time constants (60 hours) but is measured and compared in each instance. In these simulations it is assumed that a peep, on becoming symptomatic, is isolated in the school hospital: a separate building from other school activities. Hero time in this simulation is used to take account of the fact that as symptoms develop, the peep is not isolated until they are discovered. For example, a peep entering hero time overnight would not be discovered and isolated until later the next morning. Discovery, in the following simulations, occurs between the hours of 8am and 6pm.

It is also assumed that all peeps recover. The historical data were collected at daily intervals [Dudley, 1926]. The parameters for this data, analysed using the regression model discussed above, are shown in Table 1. The number of susceptible peeps within the school population was varied to establish the sensitivity of the fit to

historical data. The number of Susceptibles was varied between 38% and 82% of the population. Eight series of tests were undertaken with refinery that involved 250 simulations on each of 3 iterations. For each of the best fits found by refinery, monotony was used to obtain means and standard deviations for these instantiations.

Figure 3 shows the average number of cases as a function of the number of Susceptibles in the population along with the historical value. It was found that in all tests there were a number of simulations that did not result in infection spread. Thus the infection can be characterised by two averages: the average of all cases and the average of those cases that lead to infection spread. The proportion of non-propagating events does fluctuate but generally decreases as the number of Susceptibles in the population increase. The propagating averages behave predictably and would be expected to accord with conventional epidemiological simulation. The minimisation function is plotted in Figure 4 together with the functions formed from the deviations from the mean of r, K and τ. The results suggest that the number of Susceptibles in the population that were present in the 1920 outbreak was 67%±3%.

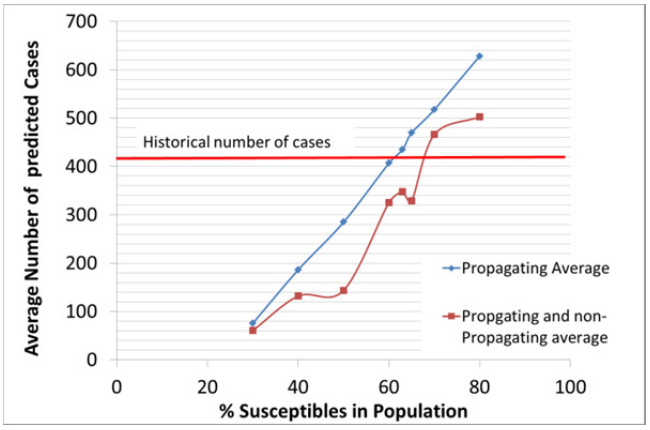

Figure 3: Variation in the number of cases, K, as a function of population susceptibility.

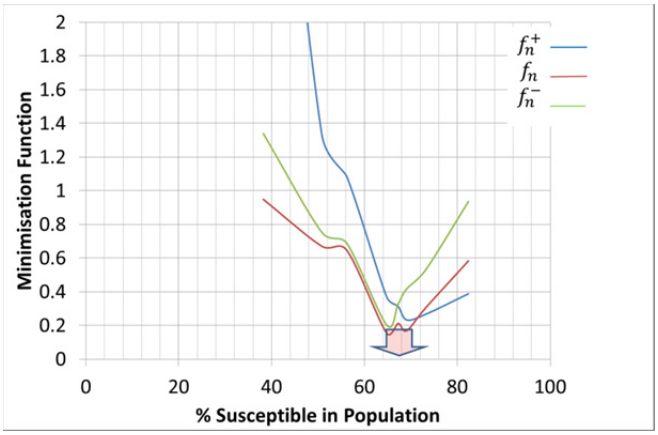

Figure 4: Minimisation functions, f_n, f_n^+, f_n^-: $f_n = \left(\Delta r_i^{\,2} + \Delta K_i^{\,2} + \Delta \tau_i^{\,2}\right)^{1/2}$; $f_n^+ = ((\Delta r_i + \mu_{ri})^2 + (\Delta K_i + \mu_{Ki})^2 + (\Delta \tau_i + \mu_{\tau i})^2)^{1/2}$; $f_n^- = ((\Delta r_i - \mu_{ri})^2 + (\Delta K_i - \mu_{Ki})^2 + (\Delta \tau_i - \mu_{\tau i})^2)^{1/2}$.

Figure 5 shows the simulations produced with monotony for a susceptible population of 67.2%; the refinery run with the closest fitness to the historical data. The simulations either did not propagate or clustered around two values of τ. In order to compare the 1920 outbreak an offset is required to be added to the historical data corresponding to the time to isolation in the simulation. However, another offset, the mean time to infection by an infected peep is required to explain the observations. In Figure 5a, the origin of the infection is one infection cycle less than assumed, where the infection cycle is the mean time to infection plus the mean time to isolation. In Figure 5b the offset corresponds to two infection cycles. The implication is that

on these examples a number of infected peeps were not recognised in the historical data. The fitness associated with the two clusters was 0.40 and 0.11 respectively. The value of $\tau = 14$ days as an estimation of the moment of the distribution curve is a much better approximation to the 1920 outbreak than 12 days used in these simulations. However the robustness of this conclusion was further tested by re-running refinery with a dataset that randomly selected the number of susceptible peeps in the population between 64% and 70% in addition to the other parameters and with the historical values of τ set at 11.4 days corresponding to the moment of the historical distribution plus one incubation period of 45 hours derived from the above data. The closest fit was expanded by monotony to obtain the mean and standard deviations for this instance.

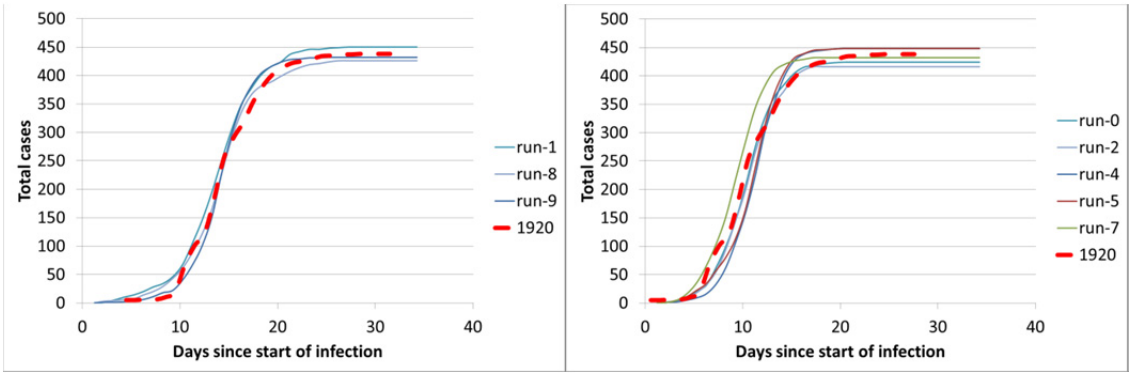

Figure 5: Total cases for 67% of the Population being susceptible: a) $\tau = 8.55$ days, b) $\tau = 12.1$days

Table 1. Infection parameters: Column 2 shows starting values for the calculations; Column 3 shows results derived from refinery and monotony.

Infection Parameter	Initial Values	Refined Values: 67.2% Susceptibles
Latent time (hours)	20 ± 4.9	20.1 ± 4.8
Asymptomatic time (hours)	15 ± 3.3	15.2 ± 3.2
Hero Time (hours)	25 ± 3.3	25.0 ± 3.2
Symptomatic Time (hours)	50 ± 0.5	50.0 ± 0.34
Immune time (hours)	2400 ± 0.5	2400 ± 0.5
Infection Chance (per contact hour)	0.0005 ± 0.00018	0.000535 ± 0.000171
Recovery Chance	100%	100%
$r \equiv 1/zR_o$	0.48	0.53
K, number of infected people	438	436
τ, moment of the cumulative function, P(t).	12.0 days	12.1 days
Attack rate, α	45%	46%
Incubation Time – time to isolation.	60 hours	45 hours
Mean time to cause infection	—	31 hours

The result confirms that the closest fit occurs with a susceptible population of about 67 ± 2 %, similar to the previous result. There were several runs produced by monotony, shown in Figure 1a, that closely matched the outbreak. An analysis of variance of the infection parameters between susceptible peeps that were involved in infection spread and those that were not showed that there was no difference in the two populations. The mean time to isolation from infection was 41.7±0.2 hours and the mean time to infection by a person already infected was 30.2±0.2 hours. Monotony produce an average reproduction number, R_0, of 3.02 with 95% confidence limits of [2.58, 3.46] for the closest instantiation. This compares favourably with a simple SIR analysis of the historical data; $R_0 = 3.08$. Taking those cases where the number of infections was within 1%

of the historical data gives $R_0 = 2.71$ with 95% confidence limits of [2.43, 3.00]. The probability of infection propagating, for a given index case, is 90% although the sample size is too low for an accurate estimate.

3 CONCLUSIONS

A method has been developed to allow stochastic studies of disease spread from microsimulation techniques. This has been demonstrated by comparison with an historical case study where social mixing and contact times were reasonably understood. The simulations show that a match is possible and gives additional information about the course of the infection and the populations involved compared to traditional techniques. The infection time constants that give the closest match to the data are not totally in agreement with those from traditional methods. Historical outbreaks of influenza suggest that the mean latent period is between 35 and 45 hours and the infectious period about 24 hours [Mathews et al., 2007], this compares with a latency of 20 hours and an effective infectious period of 22 hours (it is assumed that the data reflects the isolation that occurred at the school with an average hero time of 7 hours). Figure 1b shows that the parameters found by this method fall within a local minimum for the fitness of the solution and possibly a global one (although the range of latency needs extending to confirm this). The reason for this difference is not entirely clear but may arise from the decoupling of contact time from infection probability and the different definitions used for various time events in the literature.

ACKNOWLEDGEMENT

This research has been supported by National Science Security and Technology unit of the Department of Prime Minister and Cabinet and the ARC Linkage grants scheme.

REFERENCES

Bold, J. (2000) Greenwich, An architectural history of the royal hospital for seamen and the queens house. Yale University Press.

Connor, R. J., R. Boer, P.C. Prorok, and D. L. Weed (2000) Investigation of design and bias issues in case-control studies of cancer screening using microsimulation. *American Journal of Epidemiology 151*, 991–998.

Daley, D. J. and J. Gani (1999) Epidemic modelling: an introduction. Cambridge University Press.

Dudley, S. (1926) The spread of "droplet infection" in semi-isolated communities. *Medical Research Council Special Report.* His Majesty's Stationery Office.

Grist, R. N. (1979) Droplet infection in semi-isolated communities, pandemic influenza 1918. *British Medical Journal*, 1632–1633.

Halloran, E. M. (2001) Epidemiologic methods for the study of infectious diseases. Oxford University Press.

Keep, D., R. Bunder, I. C. Piper, and A. R. Green, (2011, August) Application of microsimulation towards modelling of behaviours in complex environments. In *Workshop on Applied Adversarial Reasoning and Risk Modeling, AAAI-11*.

Kermack, W. O., and A. G. McKendrick (1927) A contribution to the mathematical theory of epidemics, In In *Proceedings of the Royal Society of London 115*, 700–721.

Ferguson, N. M., D. A. T. Cummings, S. Cauchemez, C. Frazer, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke (2005, September). Strategies for containing an emerging influenza pandemic in south east asia. *Nature* 437(8), 209–214.

Mathews, J. D., C. T. McCaw, J. McVernon, E. S. McBryde, and J. M. McCaw (2007, November). A biological model for influenza transmission: pandemic planning: implications of asymptomatic infection and immunity. *PLoS ONE*, www.plosone.org, November 2007, Issue 11, e1220.

Modis, T. (1992), *Predictions: Society's Telltale Signature Reveals the Past and Forecasts the Future*, Simon & Schuster, New York, 1992, pp 97-105.

Orcutt, G. (1957) A new type of socio-economic system. *Review of Economics and Statistics* 39(2), 116–123.

Vergu, E., Busson, H., Ezanno, P. (2010), Impact of the Infection Period Distribution on the Epidemic Spread in a Metapopulation Model, PLoS ONE, Vol 5, Issue 2, February 2010, e9371 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2829081/pdf/pone.0009371.pdf