An Ensemble Trajectory Method for Real-Time Modeling and Prediction of Unfolding Epidemics: Analysis of the 2005 Marburg Fever Outbreak in Angola

Luís M. A. Bettencourt

Abstract We propose a new methodology for the modeling and real time prediction of the course of unfolding epidemic outbreaks. The method posits a class of standard epidemic models and explores uncertainty in empirical data to set up a family of possible outbreak trajectories that span the probability distribution of models parameters and initial conditions. A genetic algorithm is used to estimate likely trajectories consistent with the data and reconstruct the probability distribution of model parameters. In this way the ensemble of trajectories allows for temporal extrapolation to produce estimates of future cases and deaths, with quantified levels of uncertainty. We apply this methodology to an outbreak of Marburg hemorrhagic fever in Angola during 2005 in order to estimate disease epidemiological parameters and assess the effects of interventions. Data for cases and deaths was compiled from World Health Organization as the epidemic unfolded. We describe the outbreak through a standard epidemic model used in the past for Ebola, a closely related viral pathogen. The application of our method allows us to make quantitative prognostics as the outbreak unfolds for the expected time to the end of the epidemic and final numbers of cases and fatalities, which were eventually confirmed. We provided a real time analysis of the effects of intervention and possible under reporting and place bounds on population movements necessary to guarantee that the epidemic did not regain momentum.

Keywords Epidemic models · Real time estimation · Marburg-like viruses · Measurements epidemiologic · Projections and predictions

1 Introduction

Over the last few years mathematical epidemiology [1, 4] has taken an increasing interest in the quantitative study and prediction of unfolding epidemic outbreaks [2, 3, 5, 6, 21]. This is both motivated by the spectacular progress in information

L.M.A. Bettencourt (✉)
Los Alamos National Laboratory, Theoretical Division, MS B284, Los Alamos, NM 87545, USA;
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
e-mail: lmbett@lanl.gov

G. Chowell et al. (eds.), *Mathematical and Statistical Estimation Approaches in Epidemiology*, DOI 10.1007/978-90-481-2313-1_7,

technologies, which allow for the spread of epidemiological information worldwide in real time, but also to the increased monitoring of emerging infectious diseases [9, 12, 14, 22, 24–26] such as H5N1 influenza, as well as of potentially engineered biological threats [10].

The well-established tools of mathematical epidemiology built primarily for a posteriori analysis of outbreaks [1, 4] are however in several respects inadequate to measure and predict the course of unfolding epidemics. The main challenge arises from the necessary confrontation of model predictions to future data, which must be probabilistic.

Standard epidemic models, such as SIR or SEIR [1, 4], are deterministic and make a prediction for the average number of cases or deaths incurred during an outbreak. It is expected that data for large outbreaks is representative of that mean and a trajectory that fits well these data in terms of a goodness of fit measure is well accepted as the canonical procedure to estimate average epidemiological parameters.

The situation is murkier when outbreaks are small or, more to the point, when predictions from the models are to be confronted with new observations. Then the probabilistic nature of contagion becomes manifest in that no number of actual cases or deaths will usually match the predicted mean value. Thus to assess whether a model is representative of the epidemic under way it is necessary to add to this type of prediction a measure of quantified uncertainty [2, 3], e.g. in the form of a confidence interval. At that level of confidence we can then reject a model if future predictions fall outside the predicted interval (through a simple p-test), or otherwise accept the model as predictive.

This article introduces a methodology to do just this. It starts from the standard mean field models of epidemics and takes them, as specified by their initial conditions and parameter values, which we collectively denote Γ, as a possible trajectory of the outbreak. Many such trajectories are proposed via a stochastic update rule (a variant genetic algorithm) and weighted in terms of their agreement with the data at a prescribed level of uncertainty. This allows us in turn to reconstruct a probability distribution on Γ and estimate epidemiological parameters and any of their correlations with quantified uncertainty.

The remaining of this paper introduces the mathematical ensemble trajectory method and the associated estimation procedure, and then proceeds to apply it to an outbreak of a poorly known disease: Marburg hemorrhagic fever in 2005 in Angola, for which it was developed. The method made early accurate predictions of the final toll of the epidemic and its termination time and revealed erroneous trends in late reporting.

2 Uncertainty Quantification and Model Parameter Estimation

In this section we give a general description of the stochastic parameter estimation procedure. We start from the observation that simple (homogeneous mixing) population models, cannot be expected to give perfect descriptions of any actual data set. This always results in a minimum level of discrepancy between the best

model output and the data. We parameterize this discrepancy by the absolute value deviation between the best model prediction and each data point, per point. This is called the least deviation per datum (ldpd)

$$ldpp(\Gamma) = \frac{1}{N_X N_O} \sum_{i=1}^{N_X} \sum_{j=1}^{N_O} \left| X_i^M(t_j) - X_i^O(t_j) \right| \tag{A1}$$

where $X_i^M(t_j)$ is the ith state variable (e.g. deaths D, or number of cases C, see below), as an output of the model (a function of a parameter set Γ) at observation time t_j. N_X is the number of variables constrained by data and N_O is the number of observation points.

This measure allows us to discuss and compare how good models are at describing a specific data set, i.e. their goodness of fit. Secondly, we expect in general that data contain errors, e.g. due to under reporting, false positives, accounting errors, *etc*. An allowable level of uncertainty in the data will then translate into an ensemble of acceptable model of solutions or trajectories, which correspond in turn to a set of initial conditions and model parameters, which we write $\{\Gamma\}$. Each Γ in this set can then be weighted by their goodness of fit in a way that generates an estimate of the probability distribution function for the ensemble of model parameters that is compatible with the data. As a whole this is a stochastic optimization problem (see, e.g. [19] for a general discussion). Based on this idea we perform an estimation of the joint parameter distribution of model parameters $P(\Gamma)$, conditional on a set of allowable deviations per datum.

To be more specific we write that the unknown *exact* data point $X^E(t_j)$, can be expressed in terms of the observed datum $X^O(t_j)$ and an error $\xi(t_j)$ as

$$X^E(t_j) = X^O(t_j) + \xi(t_j). \tag{1}$$

The error $\xi(t_j)$ is only known statistically so that in order to proceed we need to specify a model for ξ. Because we expect the variance of the error to be bounded we assumed a Gaussian distribution for ξ, such that

$$P[\xi(t_j)] = P[X^E(t_j) - X^O(t_j)] \propto \exp\left[-\frac{\xi^2(t_j)}{2\sigma^2(t_j)} \right], \tag{2}$$

where the standard deviation $\sigma(t_j)$ parameterizes the allowed discrepancy between model outputs and data and is to be specified through general expectations on the data.

This expectation for the errors defines implicitly an objective function that can be minimized to produce optimal parameter estimates through a search procedure. For example, for each model realization in terms of a set of parameters in

a SEIR model Γ=[S(t_0), E(t_0), I(t_0), D(t_0), R(t_0), β,ε,γ,p] we take this function to be

$$A(\Gamma) = \frac{1}{N_X N_O} \sum_{i=1}^{N_X} \sum_{j=1}^{N_O} \frac{\left| X_i^M(t_j) - X_i^O(t_j) \right|^2}{2\sigma^2(t_j)}, \tag{3}$$

which is an implicit function of Γ. If the model could generate exact results we could then make the natural association $X^E(t_j) \rightarrow X^M(t_j)$. This is usually not the case, since a residual minimal deviation always persists, the minimum *ldpd*. To account for this we normalize this function to zero by taking H(Γ) = A(Γ) – A(best Γ), i.e. by subtracting the minimal value of A(Γ), obtained for the best parameter set.

Given this choice of H we can produce, in analogy with standard procedures in statistical physics, a joint probability distribution for model parameters. Since we only have expectations on A(Γ) (and not higher moments A^2, A^3, etc) the maximum entropy distribution in

$$P[\Gamma|\{X_i^O\}] \propto e^{-H}. \tag{4}$$

We can now see how this distribution can be reconstituted from sampling many realizations of the model in terms of different Γ. Note that the probability of each trajectory w(Γ) is

$$w(\Gamma) = \tfrac{1}{N_w} e^{-H(\Gamma)}, \; N_w = Tr[w_S]. \tag{5}$$

Then $P[\Gamma|\{X_i^O\}]$ can be estimated from many trajectories N_t as

$$P[\Gamma|\{X_i^O\}] \sim \sum_{i=1}^{N_t} \delta(\Gamma - \Gamma_i) w(\Gamma_i) \tag{6}$$

Figure 1 illustrates an ensemble of trajectories with variable degree of goodness of fit; trajectories with large deviations to the data points are exponentially suppressed in their contributions to the parameter distribution.

This joint probability distribution can then be used to compute any moment of any set of parameters in Γ,

$$\langle F \rangle = Tr\left[P[\Gamma \left| \{X_i^O\} \right] F(\Gamma) \right] \rightarrow \sum_{i=1}^{N_s} F(s) w(\Gamma_i) \tag{7}$$

including single parameter distribution functions, and cross-parameter correlations, such as covariances, but also any higher moments. The prediction of future observations can now be obtained by convolving the model with the parameter probability distribution estimated to that point as

$$P[X_{i+1}^O] = Tr_\Gamma \left[P[X_{i+1}^O|\Gamma, \{X_i^O\}] P[\Gamma|\{X_i^O\}] \right], \tag{8}$$

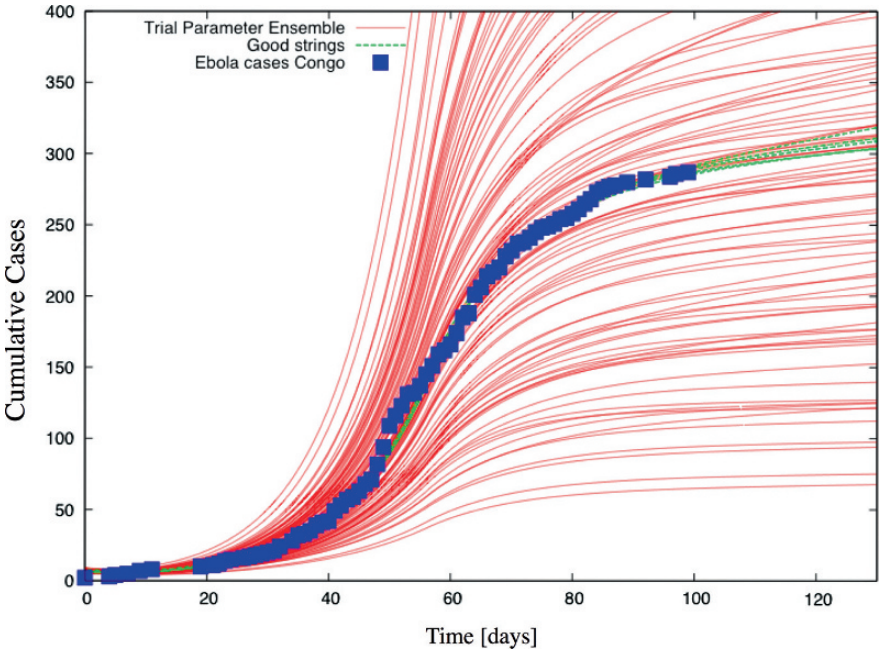

Fig. 1 Example of SEIR trajectories with varying degrees of goodness of fit for data for an outbreak of Ebola (data courtesy of Gerardo Chowell). Green trajectories fit the data (*blue squares*) well

where $P[X^O_{i+1}|\Gamma, \{X^O_i\}]$ is the model taken for a specific trajectory specified by a particular Γ.

In practice the estimation procedure via trajectories, each corresponding to a parameter set, is potentially difficult because we are dealing with an inverse problem in which, given a trial set of parameters, comparison with the data is performed only after the non-linear model dynamical equations have been solved. Fortunately for models that consist of small numbers of ordinary differential equations the computational effort is relatively trivial on a modern computer.

In every case discussed below, we used an ensemble of trial solutions, from which we select a number of best sets Γ, according to a standard Monte Carlo procedure, weighted by Eq. (5), to generate the next generation of the ensemble. In order to do this we introduce a mutation implemented in terms of random Gaussian noise around the previous best parameter set. This mutation, followed by the selection of minima, yields an effective downhill search method, capable of exploring large regions of parameter space. It also creates as a byproduct an ensemble of good strings with small deviations to the data. For small enough deviations from the best string we can sample parameter space in an unbiased manner. It is this ensemble, and its best string, that is then used to estimate Eq. (6). Results given in the manuscript involve ensembles with several million realizations and a choice of σ, common to all data points, corresponding the 20% of the *ldpd*. The standard deviation σ can be

made to vary from point to point if more information about the quality of the datum is available. In this sense the procedure is able to incorporate variable expectations of uncertainty as the data are collected.

3 Real Time Analysis of Outbreak of Marburg Fever in Angola

We now proceed to describe the application of the method to estimate in real time the course of an outbreak of the rare Marburg hemorrhagic fever in Uige, Angola during 2005. This example posed many of the challenges that lead to the development of the present methodology. The pathogen is rare and its epidemiological parameters were largely unknown beyond the observation of the apparent incubation time and time from incidence of symptoms to death in a handful of cases. The mortality was also extremely high and took the lives of many of the medical care providers that intervened early in this remote African region. Because the disease affected disproportionally children under five the implementation of isolation control measures was also extremely difficult, and their results uncertain at the time according to reports by the World Health Organization and journalists on the ground. Nevertheless even with very scant information, model predictions were accurate from one case report to the next and detected successfully the over-counting of cases and deaths that characterized late stage reports.

The current work uses data from WHO reports, freely available online or via email, together with basic epidemiological modeling to generate a characterization and outlook for the outbreak. As sparse as the data are, we hoped that our results would help quantify the progression of the disease and assess the efficacy of intervention efforts necessary to stop the epidemic. Section 3.1 gives general background information on the disease and the anatomy of the outbreak as far as it was reported at the time in medical journals and the general media. Section 3.2 describes the specific model and parameter estimation procedure. Section 3.3 analyses the scenarios of progression for the disease, taking into account the data points and some qualitative information in WHO reports. In this way we were able to estimate the effect of the interventions started shortly after March 23, in lowering contact rates, and discuss here the effects of under reporting and place bounds on population movement restrictions, so as not to reignite the epidemic. We also estimated the time horizon at which the spread would cease, as well as the final number of cases and fatalities.

3.1 Brief Anatomy of the Outbreak

The 2005 outbreak of Marburg hemorrhagic fever in Angola [13, 15–18, 27] has highlighted the direst need for fast and creative intervention in the face of the most severe infrastructure constraints imaginable. Intervention measures, which are the

only way to stop the progression of the disease in the absence of a cure, were met with significant levels of noncompliance from the population. Due to the extremely high mortality and lack of adequate preparation initially, health workers took a large toll of the early deaths, eroding confidence in their effectiveness. Furthermore the viral strand attacked primarily children under five, making it extremely difficult for families to entrust them to the healthcare system under the knowledge of very probable fatality.

Under these circumstances it was paramount to provide the best quantitative guidance and prognosis for the outbreak in real time so that limited resources can be allocated optimally. This is now starting to be possible, thanks to several outbreak surveillance and news systems, provided by the World Health Organization (WHO) [27], Pro-Med mail [15], CDC [17] and others. We used these reports to generate a data series, analyze the outbreak, and at each time elaborate scenarios for the future course of the outbreak. These can also be used to help gauge the effectiveness of the current levels of intervention and establish quantitative goals for new and/or increased measures aiming at stopping the epidemic.

The 2005 outbreak of the Marburg fever in Angola was uncommon in several respects [13, 15, 20, 27]. It was the largest of the disease to date in a general population, it had an extremely high case fatality rate (88%, compared with 23% and 70% in previous smaller outbreaks) and attacked disproportionately children under five (75% of the cases). Marburg fever symptoms in their earliest stages are non-specific. The condition can be easily confused with other more common endemic diseases in the region such as malaria, yellow fever and typhoid fever, an issue that lead to biases in reporting, especially once awareness was raised after the identification of several hundred cases and deaths. Estimates for several of the outbreak's relevant rates are [27], an incubation period of about 3–9 days, a time to death (2005 outbreak) of 3–7 days after onset of symptoms, and a high proportion of cases develop hemorrhagic symptoms within 5–7 days.

The Marburg virus is a member of the family *Filoviridae*, which also includes Ebola. Marburg however is much rarer. The reservoir of the disease remains unknown (some clues point to bats or other cave dwelling animals [11, 20]). Primates can carry the virus but also contract the disease and manifest symptoms.

Uíge is a tropical province in the interior North West of Angola, bordering the Democratic Republic of Congo. The total population of the province is estimated at about half a million people and is mostly rural. In 2005 Uíge's province two largest cities were Uíge, with about 170,000 people, and Negage with about 25,000 people. These cities' hospitals serve most of the province's population. The population of Angola is young (43.5% under 14) and with high fertility rate (6.33 children/woman), creating conditions for very high and effective transmission of the Marburg virus.

Intervention efforts by the Government and the World Health Organization started in earnest in March 23, 2005 (judging from WHO reports), a few days after the identification of the virus by the US Centers for Disease Control, Special Pathogens Branch [17, 27]. Significant efforts were also developed by non-governmental organizations such as *Medicins sans Frontiers* (Belgium, France,

Holland and Spain). Additional efforts by other international organizations are described in the WHO outbreak news reports, especially that of March 29 [27].

3.2 Homogeneously Mixing SEIR Population Model

We use a simple modification of the standard SEIR epidemic model, in order to account for the high mortality rate of the present outbreak. The model applies to homogeneously mixing population and thus does not distinguish individuals by e.g. age, a factor that is important in the current outbreak. The data necessary to draw such distinctions, if it exists at all, is not available in the public domain. Because most cases have occurred in Uíge, or are thought to have originated through contagion incurred there, we will take the total number of cases and the total number of fatalities as the targets for parameter estimation.

The SEIR model [1, 4] has been shown to describe well the outbreak dynamics of the related Ebola virus [7]. Specifically our model is:

$$\begin{aligned} \frac{dS}{dt} &= -\beta\frac{SI}{N}, \qquad \frac{dE}{dt} = \beta\frac{SI}{N} - \varepsilon E, \\ \frac{dI}{dt} &= \varepsilon E - \gamma I, \quad \frac{dD}{dt} = p\gamma I, \ \frac{dR}{dt} = (1-p)\gamma I. \end{aligned} \tag{9}$$

Here, as usual, S(t) are the number of susceptibles at time t, E(t) the number of exposed, which naturally progress to manifest the disease as Infective I(t). D(t) is the number of fatalities at time t, whereas R(t) is the number of recovered.

With these choices the population, summed over all classes, is fixed. The total number of cases, tallied at time t, is the sum of the presently infected, deceased and recovered. The incubation time is parameterized by ε^{-1}, β is the contact rate, which is the product of the (assumed independent) probability of a contact between an infected and a susceptible and the effectiveness of that contact. Lowering the value of β is the target of intervention [7]. The mean time spent in the infective class is γ^{-1}, after which an individual transits to the recovered class with probability (1-p) and dies with complementary probability p.

The set of parameters $\Gamma=\{S(t_0), E(t_0), I(t_0), D(t_0), R(t_0), \beta,\varepsilon,\gamma,p\}$, i.e. the initial conditions for each of the state variables and the dynamical parameters, is the target of our estimation procedure, as described in Section 2.

Parameter estimations will be bound within intervals dictated by knowledge of the outbreak [27]. These intervals are summarized in Table 1 below:

3.3 Parameter Estimation and Outbreak Prediction

We started tracking the outbreak in the beginning of April shortly after it was first identified on March 23. Our first predictions were made on April 26. At that time there were two possible viable scenarios for the history of the outbreak: one where it had started soon before March 23 (which we will call the simplest scenario)

Table 1 Model parameters and their allowed ranges (see text). Some of the ranges are not known and absolute maxima or minima are used

Name	Symbol	Minimum value	Maximum value
Initial Susceptibles	$S(t_0)$	10	500,000
Initial Exposed	$E(t_0)$	0	500
Initial Infective	$I(t_0)$	0	100
Initial Deceased	$D(t_0)$	0	500
Initial Recovered	$R(t_0)$	0	500
Contact rate	β	0	10
Incubation time	ε^{-1}	3 days	9 days
Lifetime infective	γ^{-1}	3 days	7 days
Case mortality	p	0	0.95

and another where it would have started sometime during October 2004. The latter was supported by retrospective analysis [27], and was eventually confirmed by our estimation procedure. We proceed to tell a brief history of our prognosis as it happened.

3.3.1 April 27: Simplest Scenario

In the simplest scenario we constrain the model by the estimated number of cases and deaths as reported by WHO, without any other further constraints. Below we consider the fact that the epidemic is though to have started in October 2004 as an additional qualitative constraint. The best fit trajectories for cases and fatalities are shown in Fig. 1, together with the data points, while parameters are displayed in Table 2.

These estimates predict an incubation time and lifetime of the infective state to be on the shorter end of their allowed ranges and mortality at the higher end. The contact rate is high leading to a large basic reproductive number, which measures the expected number of new cases caused by the introduction of an infective individual in a population of susceptibles. Given the population conditions, the high infant mortality due to the disease, and cultural practices of care for the ill and deceased we believe these numbers could not be excluded.

Table 2 Estimates for model parameters corresponding to the trajectories of Figs. 1 and 2. The initial time t_0 for parameter estimation is arbitrarily taken to be March 1

Name	Symbol	Best fit	95% CL interval
Initial Susceptibles	$S(t_0)$	200	[190, 218]
Initial Exposed	$E(t_0)$	0.2	[0.1, 0.3]
Initial Infective	$I(t_0)$	0.	[0, 1]
Initial Deceased	$D(t_0)$	87.7	[84, 90]
Initial Recovered	$R(t_0)$	0	[0, 1]
Contact rate	β	1.43	[1.18, 1.56]
Incubation time	ε^{-1}	4 days	[3.5, 5]
Lifetime infective	γ^{-1}	3 days	[3, 4]
Mortality	P	0.92	[0.91, 0.93]
Basic reproductive number	$R_0=\beta/\gamma$	4.29	[3.58, 4.69]

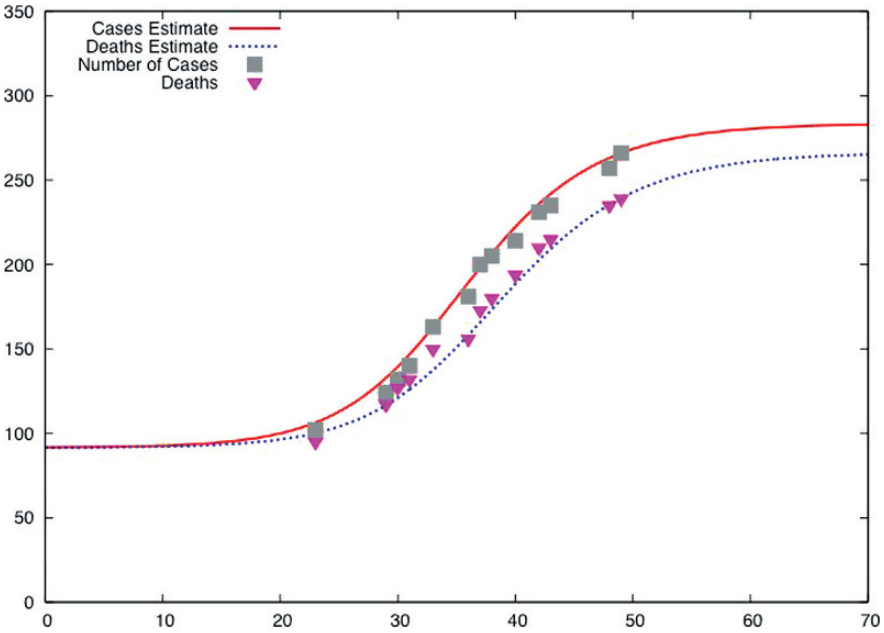

Fig. 2 Estimated best trajectories for total number of cases and deaths constrained to data from WHO surveillance reports. The least deviation per datum (ldpd) is 5.22. The curves asymptote to about 276 total cases and 261 deceased at around May 9, 2005 (the origin is the beginning of March)

This estimate predicted that the outbreak was then nearly over. The number of new infected cases was dropping in time. Its final state would be reached around May 9, with a total number of cases of 276 and 261 deaths. The upper end of the 95% confidence level intervals, shown in Fig. 2, would take these numbers up to 304 cases and 287 deaths by May 9–10. We show below that this scenario could be rejected as more data eventually came in.

Figure 3 shows the 95% confidence level intervals for number of cases and deaths. This is drawn from an ensemble of about 100,000 realizations of the model that fit the data within 20% of the best fit shown in Fig. 2.

3.3.2 Estimating Effectiveness of Intervention

The effectiveness of intervention can be assessed by allowing the contact rate β to vary in time. This strategy was used by Chowell et al. [7] to model intervention in recent Ebola epidemic outbreaks in Uganda and the Democratic Republic of Congo. The varying contact rate can be parametrized as [7]

$$\beta = \begin{cases} \beta_0, & t \leq t_{\text{int}} \\ \beta_1 + (\beta_0 - \beta_1) e^{-\kappa (t - t_{\text{int}})}, & t > t_{\text{int}} \end{cases} \tag{10}$$

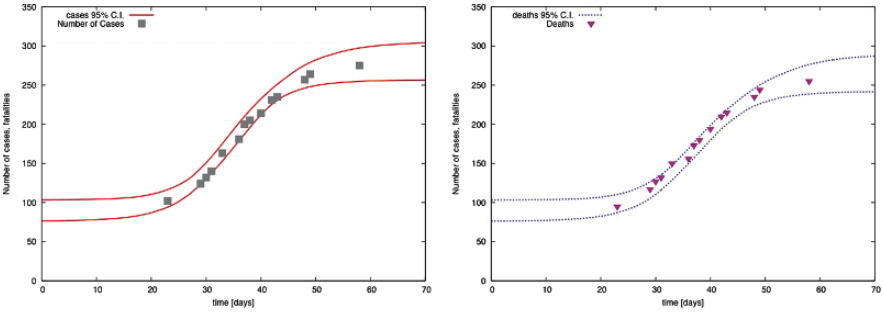

Fig. 3 95% confidence level intervals estimated from fitting the model to the data within 20% of the best fit shown in Fig. 2, for the number of cases, between *red lines* (*left*) and for number of deaths, between *blue lines* (*right*)

where t_{int} is the time at which intervention starts. κ is the time for the intervention to set in and β_0 and β_1 are the asymptotic contact rates before and after. We chose t_{int} to be March 23, when WHO reported for the first time to be "supporting efforts by the Ministry of Health in Angola to strengthen infection control in hospitals, to intensify case detection and contact tracing, and to improve public understanding of the disease and its modes of transmission" [27] (March 23 report).

We find a modest but significant change in contact rates from β_0=1.534 $\pm$ 0.013 to β_1=1.401 $\pm$ 0.010 over a period of just over 10 days. i.e. a decrease in the contact rate of about 8.7%. This gives the best fit to the data of all scenarios with ldpd=4.51. This change in contact rate highlights both the monumental efforts on the ground to contain the spread of the disease and the amply reported [8, 13] resistance they encountered, due, in large extent to the unkind characteristics of the disease.

3.3.3 Population Movements and Possible Epidemic Restart

Although the model estimated that the epidemic was then contained there can still be population movements that escape health care intervention, so that more people can enter the susceptible class. These effects can be monitored in real time via the estimation of the critical number of additional susceptibles that will cause the epidemic to regain momentum.

The simplest estimate follows from asking what number of susceptibles will reignite the growth of infected (i.e. make dI/dt>0). From Eqs. (1) this is

$$\frac{S^*}{N} = \frac{\gamma}{\beta} \tag{11}$$

For S>S* the number of new infections will grow. We can further write $S^* = S_{now} + \Delta S$ and similarly $N=N_{now} + \Delta S$, where S_{now}, N_{now} are the present numbers of susceptible and of the population participating in the epidemic (i.e. the sum of numbers over all classes) and ΔS is the critical number of additional susceptibles. Given best parameter estimates for the simple scenario on April 27 this resulted in

$\Delta S \approx 70$ individuals, which is clearly a very small fraction of the general population. We repeat this procedure in other scenarios below.

3.3.4 Impact of Possible Under Reporting and Parameter Estimation

In discussing the results of parameter estimation in this simplest scenario we found that both the case fatality rate and γ appeared at the higher end of their allowed ranges. In this section we discuss how this may be the result of case under reporting. Under reporting is probable given the remoteness of the region and the initial resistance to intervention efforts amply reported in the news [8].

To estimate the effects of under reporting we assume that number of infected reported cases I(t) is in fact a (assumed fixed) fraction λ of the real number of total cases $I^{tot}(t)$, so that

$$I(t) \to \lambda I(t) = I^{tot}(t), \qquad \lambda > 1. \tag{12}$$

We also assume that the fraction of under reporting in deaths is much smaller, so that effectively $D(t) \approx D^{tot}(t)$. We can therefore ask for the transformation in parameters that leave the dynamics of deaths invariant under the rescaling of infected. The equation become

$$\frac{dD'}{dt} = p'\gamma' I' = p'\gamma'\lambda I = \frac{dD}{dt} = p\gamma I \Rightarrow p'\gamma'\lambda = p\gamma \tag{13}$$

Thus, since $\lambda > 1$, this implies that the actual mortality is lower than estimated, and/or that the lifetime of the infectious state γ^{-1} is longer.

If we ask e.g. that $\gamma^{-1} = 5$ days and that the mortality is similar to that observed in the previous outbreak of Marburg fever in the Democratic Republic of Congo (about 70%), we would obtain

$$\lambda = \frac{p\gamma}{p'\gamma'} \cong 2.2, \tag{14}$$

suggesting that less than half of infected cases may have been reported. This was at the time most probably an overestimate. If we allow the mortality to remain above 90% then $\lambda \approx 5/3 = 1.67$, which still suggests a large fraction of unaccounted cases if the simplest scenario was to hold. This transformation has also implications for the evolution of E and S, but as these states are unconstrained by the data we shall not discuss such features here. Whether case underreporting was an explanation of the high estimates for p and γ or the modeling of the progression of the infective state was too simple in this scenario, was an issue that required more data. It was resolved by the release of the next two data points, see below.

3.3.5 April 27: Enforcing the Start of the Epidemic in October 2004

There is evidence, based on retrospective analysis [27] (March 23 report), that the epidemic started in October 2004. This section enforces such constraint in the parameter estimation procedure, analyses resulting parameter ranges and uses them to make prognoses for the development of the epidemic.

There are two caveats in performing parameter estimates under these circumstances. First, the start of the epidemic in October 2004 introduces a constraint, 4–5 months (158–121 days taking the beginning and end of the month as bounds) before the first number of cases was announced. Estimating epidemic parameters under such distant constraint is delicate and tends to lead to high sensitivity in the parameter search. As such it is intrinsically more difficult to guarantee a fair sampling of all possible solutions consistent with the data, at some error level. Second, in the very early stages of the epidemic a stochastic model is probably more appropriate than (9), which makes a number of assumptions about a homogeneously mixing population, and the applicability of averages to single instance data.

As such the results of the present estimation should be considered more susceptible to systematic error than those given above. With these caveats in mind we proceed with the estimate. Results are shown in Table 3 and in Fig. 4.

The essential qualitative consequence of enforcing that the epidemic started in October 2004 is to make the derivative in the solution for the total number of cases be positive, if small, when the virus started being tracked on March 23. Because the model (9) is monotonic in the total number of cases and deaths this necessarily generates a solution with a larger positive derivative at those first few data points. Taken at face value this constraint had two consequences: (i) it suggests that initial reported number of cases (until about March 31) were underestimates of the real numbers, (although the number of deaths is well fit by the model, and may thus not have been itself underestimated) and (ii) it led to a higher estimate – relative to the simplest scenario above – of the eventual number of cases and deaths.

Without further intervention (which was then on the way), in the absence of population movements, or any other significant external event, the epidemic was

Table 3 Estimates for model parameters corresponding to the trajectories of Figs. 4 and 5. The initial time is October 1, 2004. The estimated value of the basic reproductive number is roughly similar to that computed for recent Ebola outbreaks in Uganda and the Democratic republic of Congo [7]

Name	Symbol	Best fit	95% CL interval
Initial Susceptibles	$S(t_0)$	763	[760, 765]
Initial Exposed	$E(t_0)$	0	[0, 0.1]
Initial Infective	$I(t_0)$	0	[0, 1]
Initial Deceased	$D(t_0)$	0.5	[0, 1]
Initial Recovered	$R(t_0)$	0.35	[0, 1]
Contact rate	β	0.54	[1.18, 1.56]
Incubation time	ε^{-1}	6.5 days	[6, 7]
Lifetime infective	γ^{-1}	3 days	[3, 4]
Mortality	P	0.91	[0.90, 0.92]
Basic reproductive number	R_0	1.62	[1.60, 1.64]

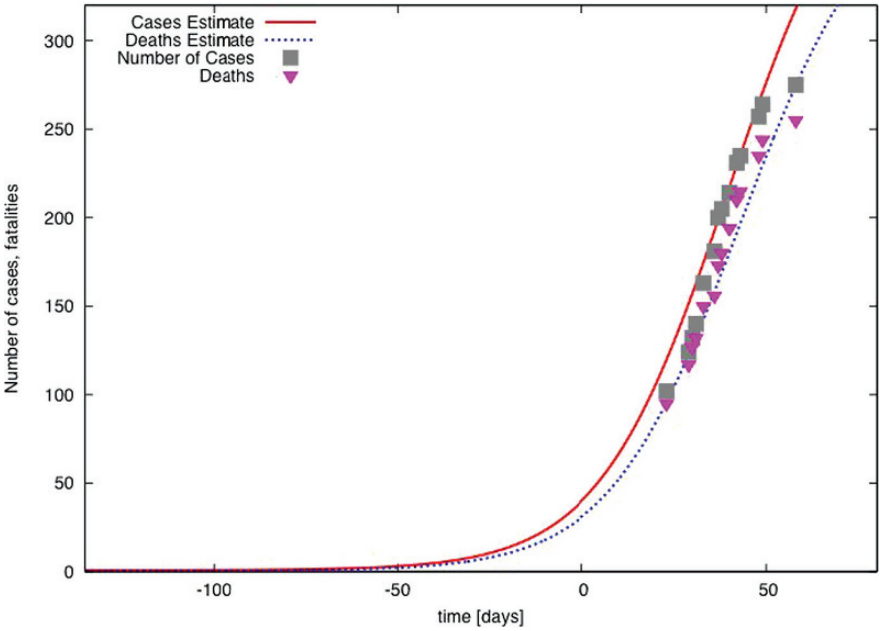

Fig. 4 Best estimated trajectories for total number of cases and deaths, under the constraint that the outbreak started October 1, 2004, without taking into account interventions. The deviation per datum is 11.15. The curves eventually asymptote to about 495 total cases, with 451 deceased by the last week of July

then expected to be extinguished only by the last week of July, with a total number of cases around 495, and 451deceased. These numbers should be taken as upper bounds. The upper end of the 95% confidence level intervals gave 519 cases and 471 deaths, whereas the lower estimated 486 cases and 440 deaths. The epidemic would

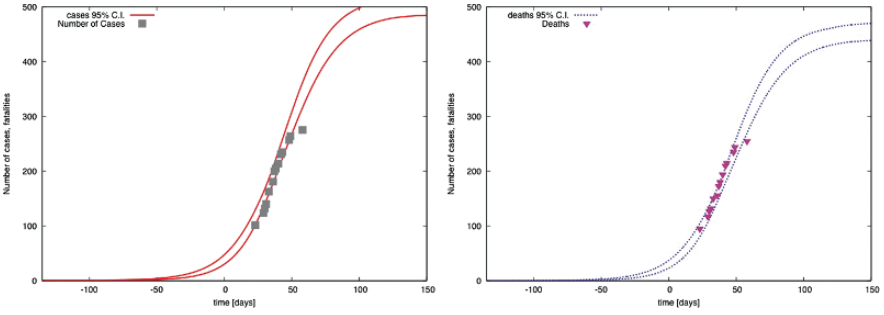

Fig. 5 The 95% confidence level intervals estimated from fitting the model to the data within 20% of the best fit shown in Fig. 4, for the number of cases, between *red lines* (*left*) and for number of deaths, between *blue lines* (*right*). The last point, reported April 27, suggests that this scenario is ruled out at 95% confidence level. This was due to interventions, see below

then need a number greater than about 180 people becoming new susceptibles, in the days after April 26, to regain growth in the number of new infectious.

Changing this constraint for the start of the epidemic from the first of October 2004 to the last generates similar, if slightly higher, estimates. This is the result of forcing the solution to be steeper, as it has to rise from zero to the case numbers reported in March over a smaller period of time. Specifically this change predicted a final number of 502 cases with 456 deaths, again by the end of July, which again would be the results in the absence of interventions. The 95% confidence level intervals are [495, 529] and [446, 480] for cases and deaths, respectively.

3.3.6 May 19: Accounting for Interventions and Distinguishing Scenarios the Beginning and End of the Outbreak

The next two data points released by the WHO in the first half of May permitted the clear distinction between the two scenarios for the beginning of the outbreak and a clear estimate of the impact of the intervention on the further development of the outbreak. Only the scenario where the outbreak started in October 2004 remained viable.

The results show that intervention – modeled by allowing β to vary according to (10), holding other parameters to the values of Table 3 – had by then managed to curb the growth rate of the outbreak, but not stop it altogether. The pool of susceptibles was estimated not to have grown over the weeks before May 9, although a small growth could not be completely excluded and was suggested by news reports, see below. The mean trajectory would eventually asymptote to an expected number of 356 total cases, with 331 deaths. This compares to the estimates (done before April 27) for 497 cases and 452 deaths, in the absence of intervention (black lines, Fig. 6). Intervention cut contact rates by a factor of about 40% and is estimated to have taken effect starting April 4, 2005 and taking about 12 days to be implemented (Fig. 7).

3.3.7 May 26: Statistical Anomalies and Over Reporting

Interestingly, as the outbreak seemed to be simmering down, the next few data points indicated a dramatic re-start of the epidemic. Results up to May 9 showed that intervention was curbing the growth rate of the outbreak, but had not succeeded at stopping it altogether. The new data released by the Angolan Government and WHO on May 26 showed a dramatic reversal of that trend. Many new cases have been registered: 399 from 337 a week before, accompanied by a sharp increase in the number of deaths to 355 from 311.

These numbers were statistical anomalies, lying far above the upper end of the 95% confidence for cases and deaths estimated on May 9. Thus they required a change in qualitative events on the ground. In the context of the model these new data points could only be accounted for in two very different scenarios. First, the new numbers could simply be wrong, attributing cases and deaths due to other causes to Marburg. Alternatively, the new data could indicate that a large number of

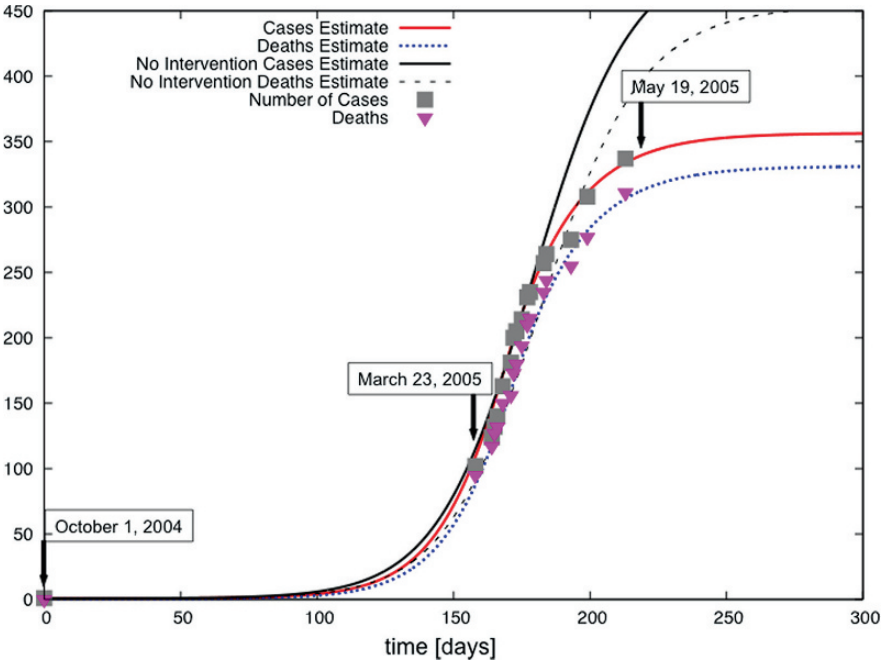

Fig. 6 Best trajectories for total number of cases and deaths, under the constraint that the outbreak started October 1, 2004 (day 0). The deviation per datum (a measure of goodness of fit) is 14.11. The color lines show the average solution under the effects of intervention and compare favorably with the trajectories estimated on April 27 (*black lines*), where this was not taken into account

new individuals (several hundred) had entered the susceptible population over the preceding few weeks and that the contact rate incurred by them had also increased, possibly up to pre-intervention levels. Clearly such a dramatic expansion of the susceptible pool should have qualitative signatures on the ground. To be true, the new data and model estimates under this scenario suggested that the epidemic threshold

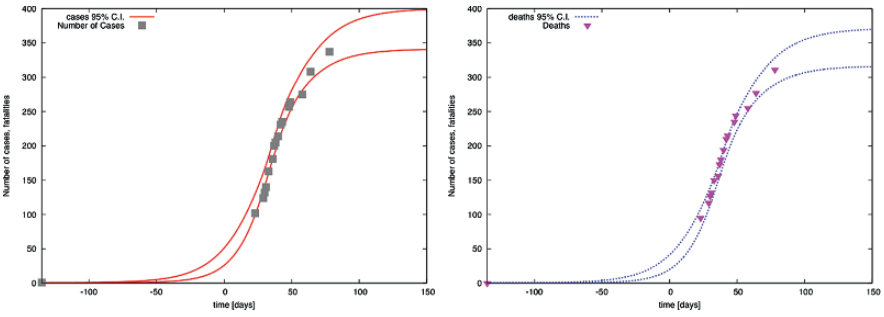

Fig. 7 95% Confidence level intervals for the number of deaths (*left*) and cases (*right*), under the constraint that the outbreak started October 1, 2004 (day 0)

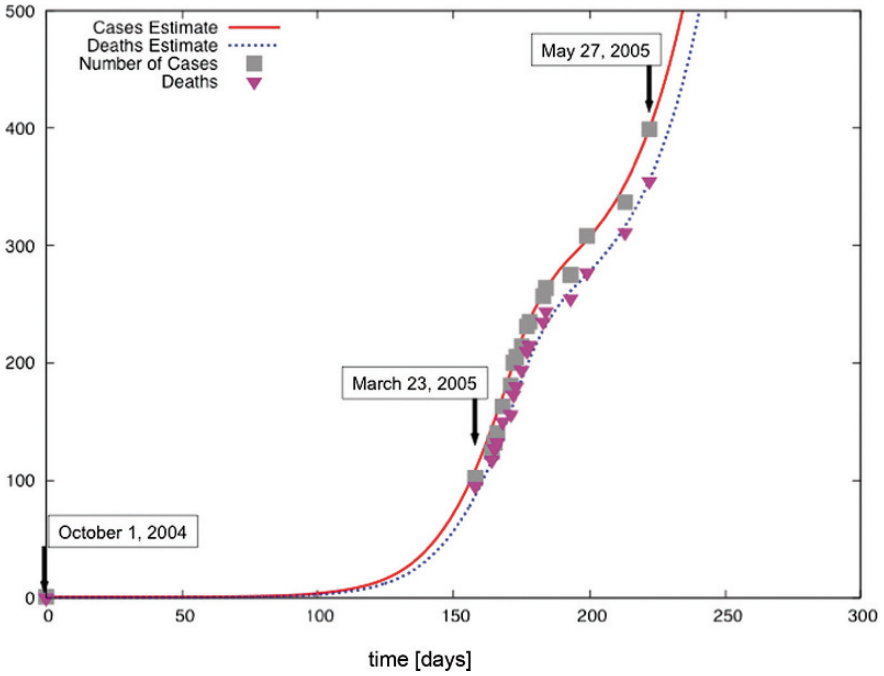

Fig. 8 Best trajectories for total number of cases and deaths, under the constraint that the outbreak started October 1, 2004 (day 0). The deviation per datum (a measure of goodness of fit) is 13.20. The new data of May 27 was an anomaly, far exceeding the upper bound of the 95% confidence level intervals for cases and deaths, estimated up to May 9. The new best fit trajectories allow β to vary upwards and require an linear inflow of people into the susceptible class at a rate of 68 persons a day

had been crossed again and that the number of new infected would subsequently grow at an accelerated pace.

In fact we could estimate that the susceptible population would have to be growing then, after the very end of April, at a rate up to 68 individuals per day. This was tantamount to an epidemic restart, visible in Fig. 8, as the average trajectories changed curvature. Needless to say, under such conditions the outlook for the development of the outbreak was rather bleak, with hundreds more cases and deaths predicted to follow.

These data points were eventually revised down, after a long hiatus in reporting between June 17 and July 13, confirming the prognosis of May 9.

3.3.8 Epilogue

The outbreak of Marburg hemorrhagic fever in Uige, Angola was officially declared over on November 7, 2005 by the Angolan Ministry of Health [23], with its last laboratory confirmed case reported July 22. The outbreak claimed a total of 329 lives out of 374 identified cases, a case fatality rate of 88%. These numbers were

correctly predicted on May 9, 2005, amid much uncertainty in WHO reports and in the news about the future development of the outbreak.

4 Discussion and Conclusions

The principal objective of the present study was to investigate the possibility of modeling in real time the spread of a new epidemic of a rare emerging disease, with very sparse data available. We have used standard outbreak reports available online from the WHO [27] and Pro-med mail [15] to construct a small data set, which we then employed to estimate epidemiological parameters and future case and death numbers with quantified uncertainty. The output of the model was used to provide guidance for the outlook of the epidemic under given qualitative scenarios and construct quantitative goals for intervention policy on the ground, creating the potential for helping optimize quantitatively severe logistical constraints.

Among other quantities our approach allows for the quantitative estimate of epidemiological parameters with quantified uncertainty, and to the projection for the total number of cases and deaths at the also predicted time for the end of the outbreak. These estimates can then be used to test qualitative scenarios about the outbreak, such as the time for the occurrence of the index case, and to quantify in real time the effects of interventions and estimate population movements. This approach, much like any general epidemiological mathematical modeling [1, 4] makes certain general simplifying assumptions about the nature of the outbreak. While these may be suspect to practitioners on the ground, it has been amply demonstrated that models retain substantial predictive power, which tends to trump projections for case numbers and deaths generated by expert opinion.

We believe that even if not perfect this type of "real time" epidemiological modeling is now feasible [2, 3, 5, 6, 21] and could become an essential tool useful in providing quantitative scenarios and targets for limited resource allocation on the ground. It should also be used to inform the scientific community and the public, as well as public health officials, of rational expectations and choices under unfolding new outbreaks.

References

1. Anderson RM, May RM (1995) Infectious Diseases of Humans. Oxford: Oxford University Press.
2. Bettencourt LMA, Ribeiro RM (2008) Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases. PLoS ONE 3(5): e2185. doi:10.1371/journal.pone.0002185
3. Bettencourt LMA, Ribeiro RM, Chowell G, Lant T, Castillo-Chavez C. Towards real time epidemiology: Data assimilation, modeling and anomaly detection of health surveillance data streams. In: Zeng D, Gotham I, Komatsu K, Lynch C, editors. Lecture Notes in Computer Science; 2007; New Brunswick, NJ: Springer-Verlag. pp. 79–90.
4. Brauer F, Castillo-Chavez C (2001) Mathematical Models in Population Biology and Epidemiology. New York: Springer-Verlag.

5. Cauchemez S, Boelle P-Y, Donnelly CA, Ferguson NM, Thomas G, et al. (2006) Real-time estimates in early detection of SARS. Emerging Infectious Diseases 12: 110–113.
6. Cauchemez S, Boelle P-Y, Thomas G, Valleron A-J (2006) Estimating in real time the efficacy of measures to control emerging communicable diseases. American Journal of Epidemiology 164: 591–597.
7. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM (2004) The reproductive number of ebola and the effects of public health measures: The cases of Congo and Uganda. Journal of Theoretical Biology 229(1):119326.
8. Denise Grady (2005) "Mysterious Viruses as Bad as They Get", New York Times April 26.
9. Fauci AS (2005) Race against time. Nature 435: 423–424.
10. Lawson AB, Kleinman K, editors (2005) Spatial and Syndromic Surveillance for Public Health. Chichester: John Wiley & Sons.
11. Leroy EM, Kumulungui B et al. (2005) Fruit bats as reservoirs of Ebola virus. Nature 438: 575–576.
12. Morens DM, Folkers GK, Fauci AS (2004) The challenge of emerging and re-emerging infectious diseases. Nature 430: 242–249.
13. Nature News http://www.nature.com/news/2005/050404/full/050404-12.html
14. Osterholm MT (2005) Preparing for the next pandemic. New England Journal of Medicine 352: 1839–1842.
15. Pro-Med mail reports http://www.promedmail.org.
16. Schou S, Hansen AK (2000) Marburg and ebola virus infections in laboratory non-human primates: A literature review. Comparative Medicine 50(2):108.
17. See Center for Disease Control briefings http://www.cdc.gov/ncidod/dvrd/spb/mnpages/dispages/marburg.htm
18. See Virginia Bioinformatics Institute Pathport site for general background http://pathport.vbi.vt.edu/pathinfo/pathogens/Marburg_virus_Info.shtml
19. Spall JC (2003) Introduction to Stochastic Search and Optimization. Hoboken NJ: Wiley.
20. Swanepoel R, Smit SB, Rollin PE, Formenty P, Leman PA, Kemp A, et al. (2007) Studies of reservoir hosts for Marburg virus. Emerging Infectious Diseases 13: 1847–1851.
21. Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of Epidemiology 160: 509–516.
22. Webby RJ, Webster RG (2003) Are we ready for pandemic influenza? Science 302: 1519–1522.
23. WHO (2005) Marburg hemorrhagic fever: Angola 2005 outbreak. Action Against Infection. 4(6).
24. Wolfe ND, Dunavan CP, Diamond J (2007) Origins of major human infectious diseases. Nature 447: 279–283.
25. Woolhouse ME, Gowtage-Sequeria S (2005) Host range and emerging and reemerging pathogens. Emerging Infectious Diseases 11: 1842–1847.
26. Woolhouse ME, Haydon DT, Antia R (2005) Emerging pathogens: the epidemiology and evolution of species jumps. Trends in Ecology & Evolution 20: 238–244.
27. World Health Organization Outbreak updates, http://www.who.int/csr/don/archive/disease/marburg_virus_disease/en/index.html