

Avoidable errors in the modeling of outbreaks of emerging pathogens, with special reference to Ebola

Aaron A. King^{1,2,3,4,*}

Matthieu Domenech de Cellès¹

Felicia M. G. Magpantay¹

Pejman Rohani^{1,2,4}

1 Department of Ecology & Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA

2 Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan, USA

3 Department of Mathematics, University of Michigan, Ann Arbor, Michigan, USA

4 Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA

* E-mail: kingaa@umich.edu

Summary

As an emergent infectious disease outbreak unfolds, public health response is reliant on information on key epidemiological quantities, such as transmission potential and serial interval. Increasingly, transmission models fit to incidence data are used to estimate these parameters and guide policy. Some widely-used modeling practices lead to potentially large errors in parameter estimates and, consequently, errors in model-based forecasts. Even more worryingly, in many such situations, confidence in parameter estimates and forecasts is itself far over-estimated, leading to the potential for large errors that mask their own presence. Fortunately, straightforward and computationally inexpensive alternatives exist that avoid these problems. Here, we first use a simulation study to demonstrate potential pitfalls of the standard practice of fitting deterministic models to cumulative incidence data. Second, we demonstrate an alternative based on stochastic models fit to raw data from the 2014 West African Ebola virus disease outbreak. We show not only that bias is thereby reduced, but that uncertainty in estimates and forecasts is better quantified and that, critically, lack of model fit is more readily diagnosed.

Introduction

The success of model-based policy in the instances of bovine spongiform encephelopathy ([Anderson *et al.*, 1996](#)) and foot-and-mouth disease ([Ferguson *et al.*, 2001](#); [Keeling *et al.*, 2001](#)) has increased expectations that forecasts based on scientifically informed models of disease transmission are a necessary tool for response to outbreaks of emerging infectious diseases. Increasingly, the expectation is that reliable forecasts will be available in real time. Recent examples in which model-based forecasts were produced within weeks of the index case include severe acute respiratory syndrome (SARS; [Lipsitch *et al.*, 2003](#); [Riley *et al.*, 2003](#)), pandemic H1N1 influenza ([Fraser *et al.*, 2009](#)), cholera in Haiti and Zimbabwe ([Tuite *et al.*, 2011](#)), Middle East respiratory syndrome (MERS; [Breban *et al.*, 2013](#)), and lately, Ebola virus disease (EBVD) in west Africa ([Fisman *et al.*, 2014](#)). In the early stages of an emerging pathogen outbreak, key unknowns include its transmission potential, the likely magnitude and timing of the epidemic peak, total outbreak size, and the durations of the incubation and infectious phases. Many of these quantities can be estimated using clinical and household transmission data, which are, by definition, rare in the early stages of such an outbreak. Much interest therefore centers on estimates of these quantities from

incidence reports that accumulate as the outbreak gathers pace. Such estimates are obtained by fitting mathematical models of disease transmission to incidence data.

As is always the case in the practice of confronting models with data, decisions must be made as to the structure of fitted models and the data to which they will be fit. Concerning the first, in view of the urgency of policy demands and paucity of information, the simplest models are, quite reasonably, typically the first to be employed. With even the simplest models, such as the classical susceptible-infected-recovered (SIR) model, the choice of data to which the model is fit can have significant implications for science and policy. Here, we explored these issues using a combination of inference on simulated data and on actual data from the ongoing EBVD outbreak. We find that some of the standard choices of model and data can lead to potentially serious errors. Moreover, regardless of the model choice, it is important to bear in mind that all model-based conclusions hinge on the ability of the model to fit the data. We demonstrate a straightforward alternative that allows model misspecification to be diagnosed and forecast uncertainty to be properly quantified.

Deterministic models fit to cumulative incidence curves: a recipe for error and overconfidence

An inexpensive and therefore common option is to formulate deterministic transmission models and fit these to data using least squares or related methods. These approaches seek parameters for which model trajectories pass as close to the data as possible. Because in such approaches the model itself is deterministic, all discrepancies between model prediction and data are effectively ascribed to measurement error. As for the data to be fit, it has been tempting to fit model trajectories to cumulative case counts. The incompatibility of this choice with the assumptions of the statistical error model has been pointed out previously (Grad *et al.*, 2012; Ma *et al.*, 2014; Towers *et al.*, 2014). In particular, the validity of the statistical estimation procedure hinges on the independence of sequential measurement errors, which is clearly violated when the data are accumulated through time. To explore the impact of this violation on inferences and projections, we performed a simulation study in which we generated data using a stochastic model, then fit the corresponding deterministic model to both raw and cumulative incidence curves. We generated 200 sets of simulated data for three different levels of measurement noise. For each data set, we estimated model parameters, including transmission potential (as quantified by the basic reproduction number, R_0), reporting probability, and observation error. Full details of the data generation and fitting procedures are given below, under Methods. The resulting parameter estimates are shown in Fig. 1.

Recognizing that quantification of uncertainty is prerequisite to reliable forecasting, we computed parameter estimate confidence intervals, and investigated their accuracy. Fig. 1A shows little evidence for bias in estimates of R_0 whether raw or cumulative incidence data are used. This result is likely non-generic: in general, violation of the independent-error assumption will result in some degree of additional bias. Fig. 1B is the corresponding plot of estimated overdispersion of measurement noise. Using the raw incidence data, one recovers the true observation variability. When fitted to cumulative data, however, the estimates display extreme bias: far less measurement noise is needed to explain the relatively smooth cumulative incidence. The data are apparently in good agreement with the model. To quantify the uncertainty in the parameter estimates, we examined the confidence intervals. The nominal 99% profile-likelihood confidence interval widths for R_0 are shown in Fig. 1C. When the model is fit to the simulated data, increasing levels of measurement error lead to increased variance in the estimates of R_0 . However, the confidence interval widths are far smaller when the cumulative data are used, superficially suggesting a higher degree of precision. This apparent precision is an illusion however, as Fig. 1D shows. This

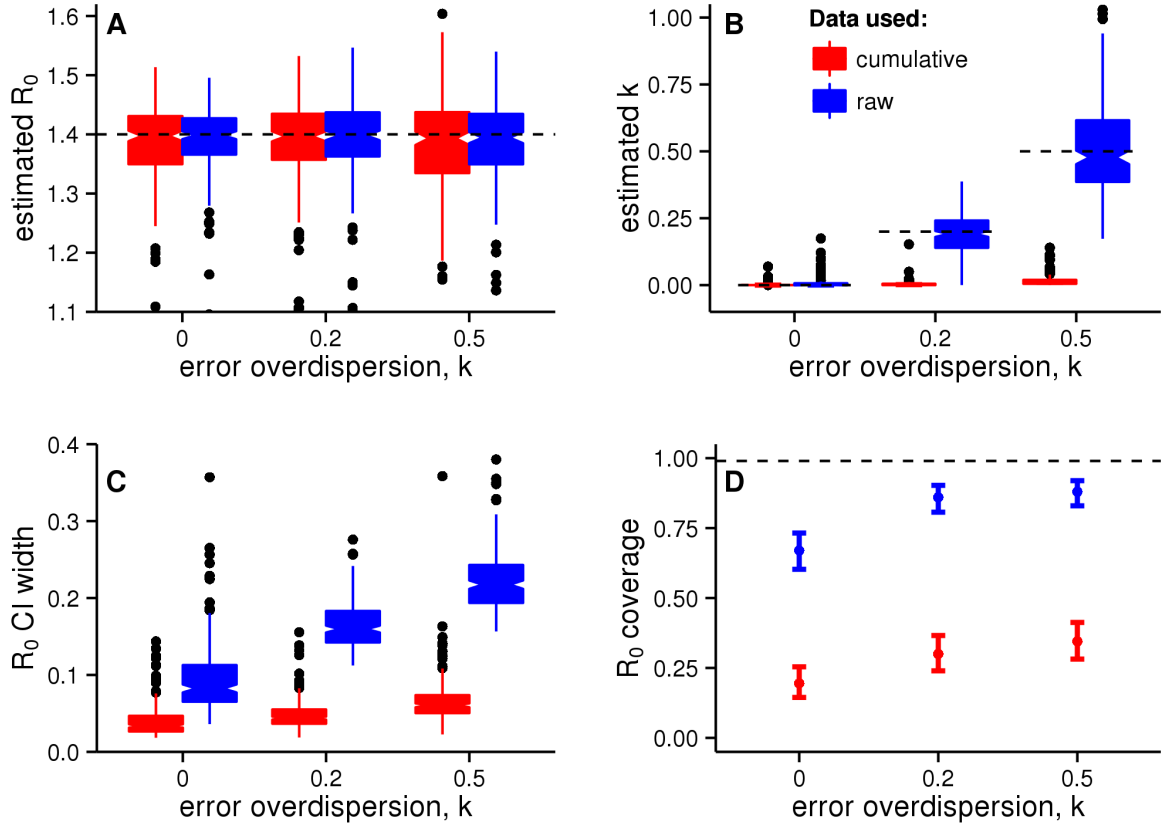


Figure 1: Results from simulation study fitting deterministic models to stochastically simulated data. 200 simulated data sets of length 39 wk were generated by the stochastic model described in the Methods section at each of three levels of the measurement error overdispersion parameter, k . The deterministic model was fit to both raw (blue) and accumulated (red) incidence data. (A) Estimates of R_0 . True value used in generating the data is shown by the dashed line. (B) Estimates of k . (C) Widths of nominal 99% profile likelihood confidence intervals (CI) for R_0 . (D) Actual coverage of the CI, i.e., probability that the true value of R_0 lay within the CI. Ideally, actual coverage would agree with nominal coverage (99%, dashed line).

figure plots the achieved coverage (probability that the true parameter value lies within the estimated confidence interval) as a function of the magnitude of measurement error and the choice of data fitted. Given that the nominal confidence level here is 99%, it is disturbing that the true coverage achieved is closer to 25% when cumulative data are used.

When a deterministic model is fit to cumulative incidence data, the net result is a potentially substantially over-optimistic estimate of precision, for three reasons. First, failure to account for the non-independence of successive measurement errors leads to an under-estimate of parameter uncertainty (Fig. 1C). Second, as seen in Fig. 1B, the variance of measurement noise will be substantially under-estimated. Finally, because the model ignores environmental and demographic stochasticity, treating the unfolding outbreak as a deterministic process, forecast uncertainty will grow unrealistically slowly with the forecast horizon. We elaborate on the last point in the Discussion.

Stochastic models fit to raw incidence data: feasible and transparent

Though stochastic models are a bit more complex (Keeling & Rohani, 2008), their inclusion of demographic and/or environmental stochastic processes allows for, on the one hand, better fits to the trends and variability in outbreak data and, on the other, improved ability to diagnose lack of model fit (He *et al.*, 2010). We formulated a stochastic version of the SEIR model as a partially observed Markov process and fit it to actual data from the ongoing West African EBVD outbreak. We estimated parameters by maximum likelihood, using sequential Monte Carlo to compute the likelihood and iterated filtering to maximize it over unknown parameters. See below, under Methods, for details. To tease apart the consequences of failing to account for stochasticity from those of improperly fitting to cumulative data, we fit the stochastic model both to actual incidence and to cumulative incidence data. In the latter case, we made the standard, erroneous, assumption of independence of successive measurement errors.

Fig. 2 shows likelihood profiles over R_0 for country-level data from Guinea, Liberia, and Sierra Leone. We also wanted to explore potential biases associated with spatial aggregation of the data. Hence, we fit our models to regional data, encompassing all reported cases from the three West African countries just mentioned. The profile likelihoods show clearly that the additional parameters of the stochastic model do not erase the identifiability of R_0 . As was the case with the deterministic model fits, little bias in point estimates of R_0 is evident when models are fit to raw vs. cumulative data. Secondly, and also in line with the lessons of Fig. 1C, estimated confidence intervals are narrower when the cumulative reports are used. The “true” parameters are, of course, unknown, but, as in the earlier example, this higher precision is probably illusory. The somewhat, but not dramatically, larger confidence intervals that come with adherence to the independent errors assumption (i.e., with the use of raw incidence data) lead to a quite substantial increase in forecast uncertainty, as we shall see. Finally, the ease with which the stochastic model was fit and likelihood profiles computed testifies to the fact that, in the case of outbreaks of emerging infectious diseases, it is not particularly difficult or time-consuming to work with stochastic models.

We took advantage of the stochastic model formulation to diagnose the fidelity of model to the data. To do so, we simulated 10 realizations of the fitted model; the results are plotted in Fig. 3. While the overall trends appear similar, the model simulations display greater variability at high frequencies than do the data. To quantify this impression, we computed the correlation between cases at weeks t and $t - 1$ (i.e., the autocorrelation function at lag 1 wk, $ACF(1)$) for both model simulations and data. For Guinea,

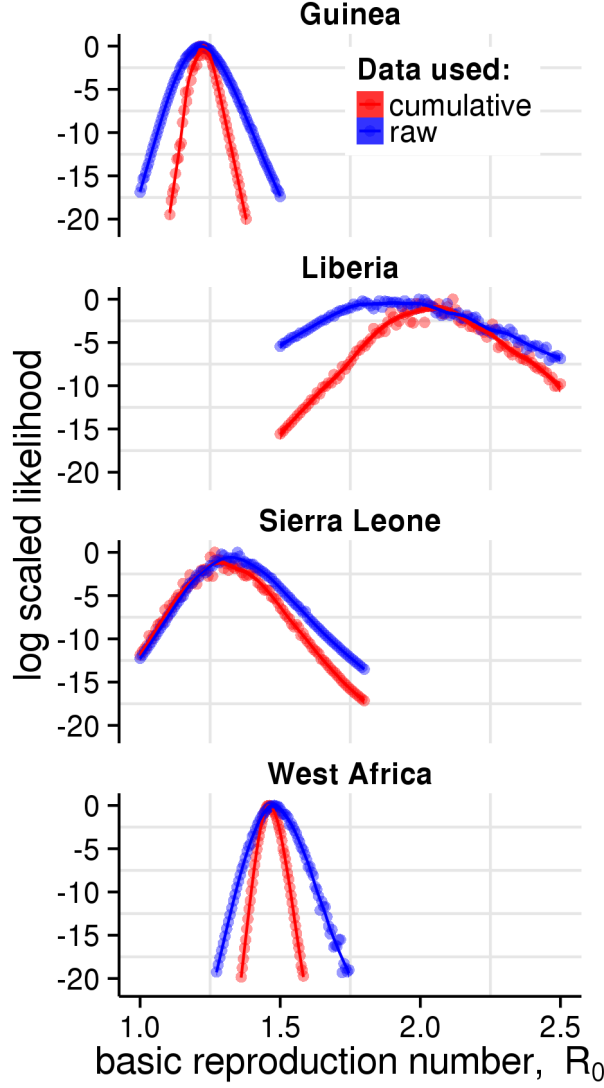


Figure 2: Likelihood profiles for R_0 based on the stochastic model using raw and cumulative incidence data. Each point represents the maximized log likelihood at each fixed value of R_0 relative to overall maximum. The maximum point of each profile is the maximum-likelihood estimate (MLE) of R_0 ; the curvature is proportional to estimated precision. While the (improper) use of cumulative data produces relatively small errors in the MLE, it does produce the illusion of high precision. The procedure for maximizing the likelihood is described under Methods, below.

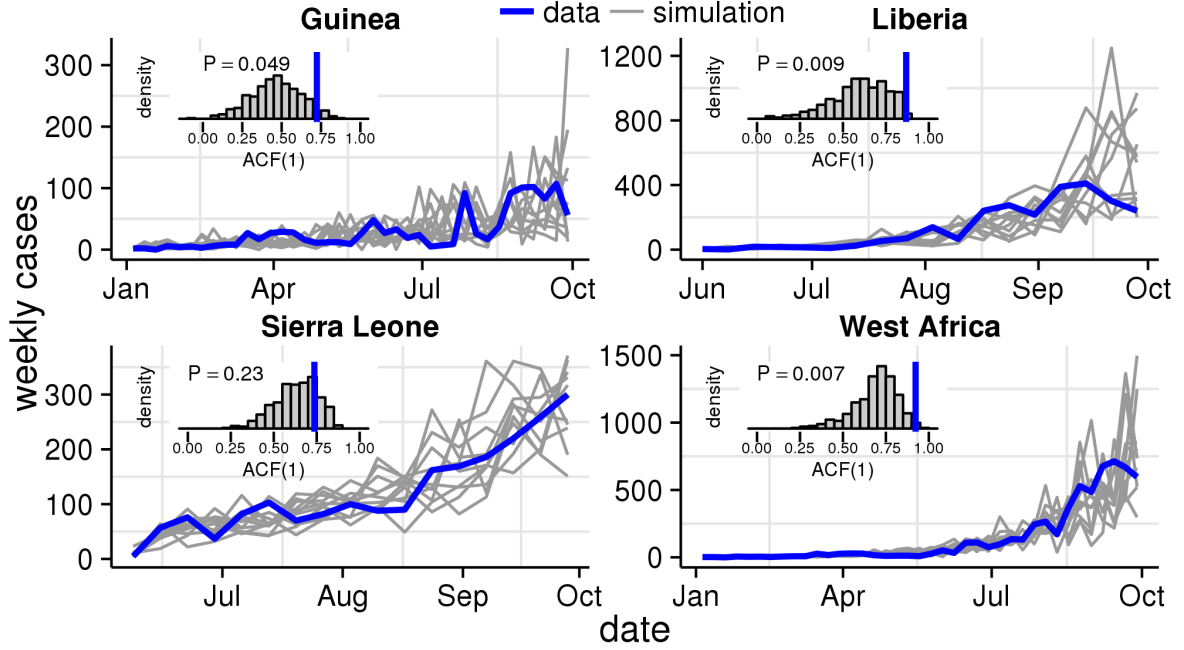


Figure 3: Model diagnostics. The time series plots show the data (blue) superimposed on 10 typical simulations from the fitted model (grey). While the overall trend is captured by the model, the simulations display more high-frequency (week-to-week) variability than does the data. The insets confirm this, showing the autocorrelation function at lag 1 week ($ACF(1)$) in the data (blue) superimposed on the distribution of $ACF(1)$ in 500 simulations (grey). For Guinea, Liberia, and the aggregated regional data (“West Africa”), the $ACF(1)$ of the data lies in the extreme right tail of the distribution, as quantified by the one-sided P -values shown.

Liberia, and the region as a whole (“West Africa”), the observed $ACF(1)$ lies in the extreme right tail of the model-simulated distribution, confirming our suspicion. For Sierra Leone, the disagreement between fitted model and data is not as poor, at least as measured by this criterion. A similar picture emerges from an analysis of the correlation structure of model residuals (see Methods). Further, the estimated reporting probabilities of 0.01 in Guinea and 0.024 in Liberia (corresponding to, respectively, 96 and 41 unrecorded infections per notified case) are at odds with current field estimates of these quantities (Bellan *et al.*, 2014). Taken together, these diagnostics caution against the interpretation of the outbreaks in Guinea and Liberia as simple instances of SEIR dynamics and call for a degree of skepticism in inferences and forecasts based on this model. On the other hand, the Sierra Leone epidemic does appear to better conform to the SEIR assumptions, at least when the data are aggregated to the country level, and has plausible parameter estimates.

Fig. 4 suggests why the present Ebola outbreak might not be adequately described by the well-mixed dynamics of the SEIR model. The erratically fluctuating mosaic of localized hotspots suggests spatial heterogeneity in transmission, at odds with the model’s assumption of mass action. As an aside, this heterogeneity hints at control measures beyond the purview of the SEIR model. While the latter might provide more or less sound guidance with respect to eventual overall magnitude of the outbreak and

associated demands for hospital beds, treatment centers, future vaccine coverage, etc., the former points to the potential efficacy of movement restrictions and spatial coordination of control measures.

Discussion

To summarize, we have here shown that the frequently adopted approach of fitting deterministic models to cumulative incidence data can lead to biased estimation of some parameters and a dramatic underestimation of the uncertainty associated with model parameters. Unsurprisingly, forecasts based on such approaches are similarly plagued by difficult-to-diagnose over-confidence as well as bias. We illustrate this using the SEIR model—in its deterministic and stochastic incarnations—fit to data from the current Sierra Leone EBVD outbreak. Emphatically, we do not here assert that the SEIR model adequately captures those features of the Sierra Leone epidemic needed to make accurate forecasts. Indeed, we have side-stepped important issues of identifiability of key parameters such as route-specific transmissibility and effective infectious period. Rather, we have erred on the side of parsimony to better focus on the issues we raise above in the context of quantities of immediate and obvious public health importance, particularly the basic reproduction number and trajectory of the ongoing outbreak. Fig. 5 shows projected incidence of EBVD in Sierra Leone under both the deterministic model fit to cumulative incidence data (in red) and the stochastic model fit to raw incidence data (in blue). The shaded ribbons indicate forecast uncertainty. In the deterministic case, the latter is due to the combined effects of estimation error and measurement noise. As we showed above, the first contribution is unrealistically low because serial autocorrelation among measurement errors have not been properly accounted for. The second contribution is also under-estimated because of the smoothing effect of data accumulation. Finally, because the model ignores all process noise, it unrealistically lacks dynamic growth of forecast uncertainty. By contrast, the stochastic model fitted to the raw incidence data show much greater levels of uncertainty. Because measurement errors have been properly accounted for, confidence intervals more accurately reflect true uncertainty in model parameters. Because the model accounts for process noise, uncertainty expands with the forecast horizon. Finally, we recall once again that, because the process noise terms can to some degree compensate for model misspecification, it was possible to diagnose the latter, thus obtaining some additional qualitative appreciation of the uncertainty due to this factor.

The increasingly high expectations placed on models as tools for public policy put an ever higher premium on the reliability of model predictions. The relentless tradeoff between timeliness and reliability has with technological advance shifted steadily in favor of more complex and realistic models. Because stochastic models with greater realism, flexibility, and transparency can be routinely and straightforwardly fit to outbreak data, there is less and less scope for older, less reliable, and more opaque methods. In particular, the practices of fitting deterministic models and fitting models to cumulative case report data are prejudicial to accuracy and can no longer be justified on pragmatic grounds. We propose the following principles to guide modeling responses to current and future infectious disease outbreaks:

1. Models should be fit to raw, disaggregated data whenever possible and never to temporally accumulated data.
2. When model assumptions, such as independence of errors, must be violated, careful checks for the effects of such violations should be performed.
3. Forecasts based on deterministic models, being by nature incapable of accurately communicating uncertainty should be avoided.

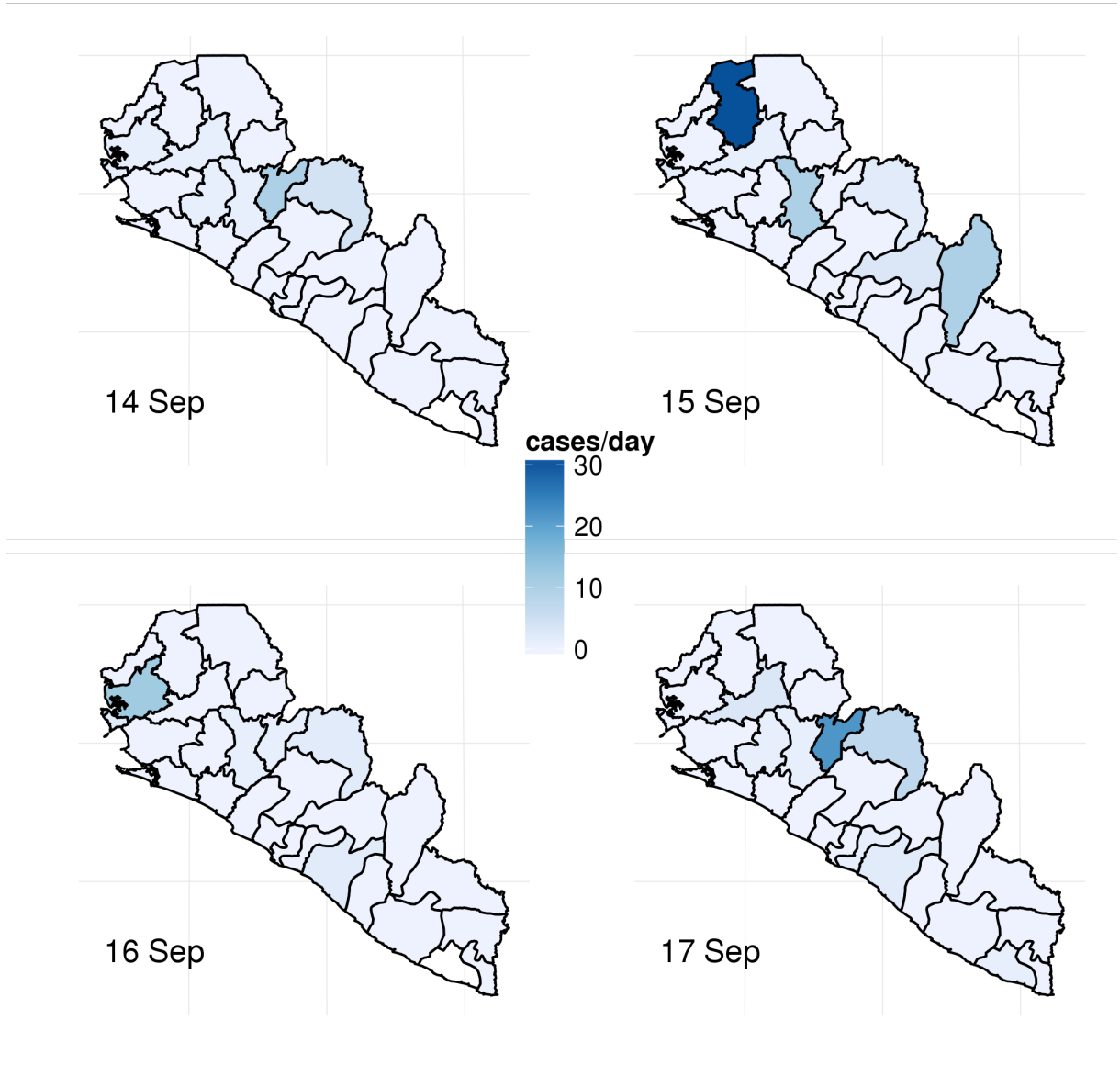


Figure 4: Four consecutive days of Ebola incidence in the republics of Liberia and Sierra Leone. In the outbreaks early stages, the spatio-temporal dynamics are highly erratic, contrary to the predictions of the well-mixed model.

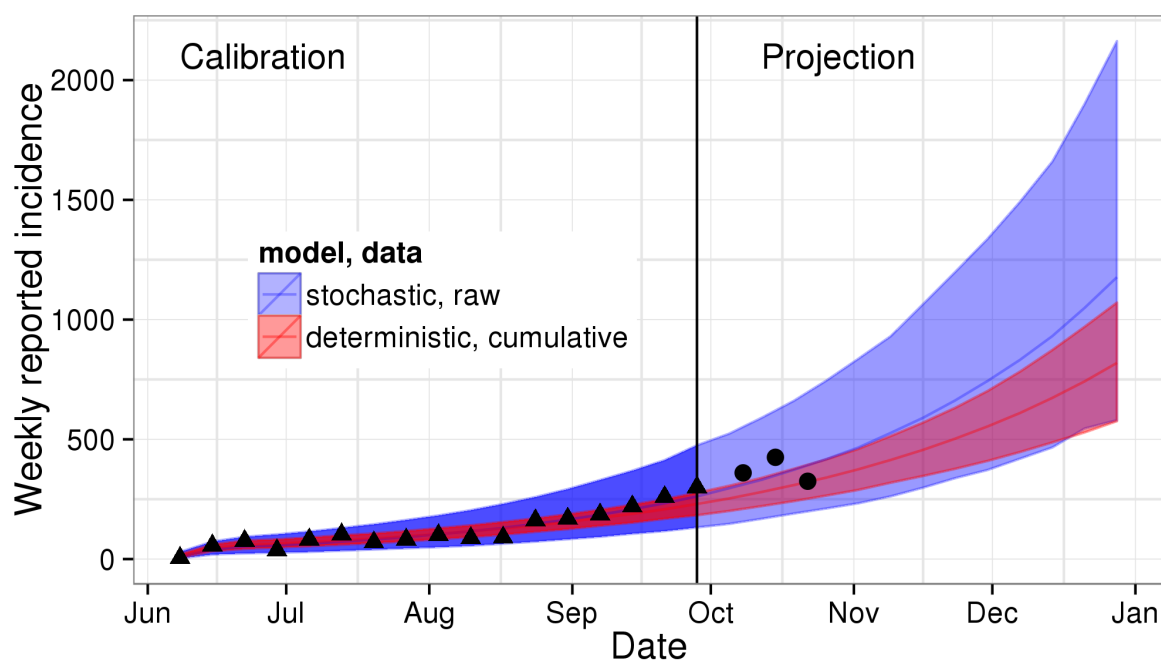


Figure 5: Forecast uncertainty for the Sierra Leone EBVD outbreak as a function of the model used and the data to which the model was fit. The red ribbon shows the median and 95% envelope of model simulations for the deterministic SEIR model fit to cumulative case reports; the blue ribbon shows the corresponding forecast envelope for the stochastic model fit to raw incidence data. The data used in model fitting are shown using black triangles; out-of-fit data are shown by black circles.

4. Stochastic models should be preferred to deterministic models in most circumstances because they afford improved accounting for real variability and increased opportunity for quantifying uncertainty and diagnosing model misspecification.

In closing, we worry that screening for lack of model fit is not a standard part of modeling protocol. At best, this represents a missed opportunity, as discrepancies between the data and off-the-shelf models may suggest effective control measures. At worst, this can lead to severely biased estimates and, worryingly, overly confident conclusions. Fortunately, effective techniques exist by which such errors can be diagnosed and avoided, even in circumstances demanding great expedition.

Methods

Data

Weekly case reports in Guinea, Liberia, and Sierra Leone were digitized from the WHO situation report dated from 1 October 2014¹ (Fig. 3). To compare our predictions to those of previous reports (Gomes *et al.*, 2014), we also aggregated those data to form a regional epidemic curve for “West Africa”. In Guinea, this outbreak was taken to have started in the week ending 5 January 2014 and in Sierra Leone in that ending 8 June 2014. In Liberia, the outbreak was notified to WHO on 31 March 2014², but few cases were reported until June; therefore, the week ending 1 June was deemed the start of the Liberian outbreak for simulation purposes. The data in Fig. 4 was downloaded from the repository maintained by C. M. Rivers³ and ultimately derived from reports by the health ministries of the republics of Guinea, Sierra Leone, and Liberia.

Model formulation

The models used were variants on the basic SEIR model model, using the method of stages to allow for a more realistic (Gamma) distribution of the incubation period (Lloyd, 2001; Wearing *et al.*, 2005). The equations of the deterministic variant are:

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{R_0\gamma SI}{N} \\
\frac{dE_1}{dt} &= \frac{R_0\gamma SI}{N} - m\alpha E_1 \\
\frac{dE_i}{dt} &= m\alpha(E_{i-1} - E_i), \quad i = 2, \dots, m \\
\frac{dI}{dt} &= m\alpha E_m - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{aligned}$$

¹<http://www.who.int/csr/disease/ebola/situation-reports/en/>

²<http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4072-ebola-virus-disease-liberia.html>

³<https://github.com/cmrvivers/ebola>

Here, R_0 represents the basic reproduction number; $1/\alpha$, the average incubation period; m , the shape parameter for the incubation period distribution; $1/\gamma$, the average infectious period; and N , the population size, assumed constant (Table S1).

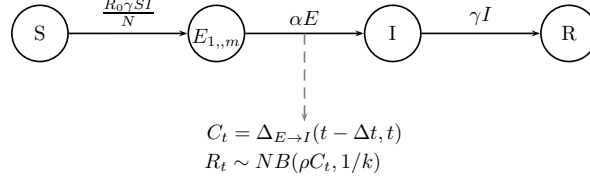


Figure 6: Schematic diagram of the transmission models used. NB = negative binomial distribution. $\Delta_{E \rightarrow I}(t - \Delta t, t)$ = number of transitions from latent to infectious class (E to I) occurring between times $t - \Delta t$ and t .

The stochastic variant was implemented as a continuous-time Markov process approximated via a multinomial modification of the τ -leap algorithm (He *et al.*, 2010) with a fixed time step $\Delta t = 10^{-2}$ wk.

To complete the model specification, we model the observation process. Between times $t - \Delta t$ and t , where Δt represents the reporting period, we write $C_t = \Delta N_{E \rightarrow I}(t - \Delta t, t)$ for the complete number of new infections during that time period. When we are fitting to cumulative case counts, we change the definition accordingly to $C_t = \Delta N_{E \rightarrow I}(0, t)$. When using either type of data, we modeled the corresponding case report, R_t , as a negative binomial: $R_t \sim NB(\rho C_t, 1/k)$ (i.e., $\mathbb{E}(R_t) = \rho C_t$, $\mathbb{V}(R_t) = \rho C_t + k \rho^2 C_t^2$), where ρ is the reporting probability and k the reporting overdispersion.

Descriptions of the methods used in the simulation study and in the model-based inferences drawn from actual data are given in the Appendix.

Acknowledgements

We thank John Drake, Andrew Park, and Robert Reiner for comments. PR and AAK are supported by the Research and Policy in Infectious Disease Dynamics program of the Science and Technology Directorate, Department of Homeland Security, the Fogarty International Center, National Institutes of Health and by MIDAS, National Institute of General Medical Sciences U54-GM111274 and U01-GM110744.

References

- Anderson, R. M., Donnelly, C. A., Ferguson, N. M., Woolhouse, M. E., Watt, C. J., Udy, H. J., MaWhinney, S., Dunstan, S. P., Southwood, T. R., Wilesmith, J. W., Ryan, J. B., Hoinville, L. J., Hillerton, J. E., Austin, A. R. & Wells, G. A., 1996 Transmission dynamics and epidemiology of BSE in British cattle, *Nature* **382**, 779–788.
- Bellán, S. E., Pulliam, J. R. C., Dushoff, J. & Meyers, L. A., 2014 Ebola control: effect of asymptomatic infection and acquired immunity, *Lancet* pp. –.

- Breban, R., Riou, J. & Fontanet, A., 2013 Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk., *Lancet* **382**, 694–699.
- Chowell, G. & Nishiura, H., 2014 Transmission dynamics and control of Ebola virus disease (EVD): a review., *BMC Medicine* **12**, 196.
- Ferguson, N. M., Donnelly, C. A. & Anderson, R. M., 2001 Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain (vol 413, pg 542, 2001), *Nature* **414**, 329–329.
- Fisman, D., Khoo, E. & Tuite, A., 2014 Early epidemic dynamics of the West African 2014 Ebola outbreak: Estimates derived with a simple two-parameter model, *PLoS Currents Outbreaks* .
- Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela, I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C. & W. H. O. R. P. A. C., 2009 Pandemic potential of a strain of influenza A (H1N1): early findings, *Science* **324**, 1557–1561.
- Gomes, M., y Piontti, A. P., Rossi, L., Chao, D., Longini, I., Halloran, M. E. & Vespignani, A., 2014 Assessing the international spreading risk associated with the 2014 West African Ebola outbreak, *PLoS Currents Outbreaks* .
- Grad, Y. H., Miller, J. C. & Lipsitch, M., 2012 Cholera modeling: challenges to quantitative analysis and predicting the impact of interventions, *Epidemiology* **23**, 523–530.
- He, D., Ionides, E. L. & King, A. A., 2010 Plug-and-play inference for disease dynamics: measles in large and small populations as a case study, *Journal of The Royal Society Interface* **7**, 271–283.
- Ionides, E. L., Bhadra, A., Atchadé, Y. & King, A. A., 2011 Iterated filtering, *Annals of Statistics* **39**, 1776–1802.
- Ionides, E. L., Bretó, C. & King, A. A., 2006 Inference for nonlinear dynamical systems, *Proceedings of the National Academy of Sciences of the U.S.A.* **103**, 18438–18443.
- Keeling, M. & Rohani, P., 2008 *Modeling Infectious Diseases in Humans and Animals*, Princeton: Princeton University Press, ISBN 0691116172.
- Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J. & Grenfell, B. T., 2001 Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape, *Science* **294**, 813–817.
- King, A. A., Ionides, E. L., Bretó, C. M., Ellner, S., Kendall, B., Wearing, H., Ferrari, M. J., Lavine, M. & Reuman, D. C., 2010 *pomp: Statistical inference for partially observed Markov processes (R package)*.
- Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H., Fisman, D. & Murray, M., 2003 Transmission dynamics and control of severe acute respiratory syndrome, *Science* **300**, 1966–1970.
- Lloyd, A. L., 2001 Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods, *Proceedings of the Royal Society of London. Series B* **268**, 985–993.
- Ma, J., Dushoff, J., Bolker, B. M. & Earn, D. J. D., 2014 Estimating initial epidemic growth rates, *Bulletin of Mathematical Biology* **76**, 245–260.

- Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. & Timmer, J., 2009 Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, *Bioinformatics* **25**, 1923–9.
- Riley, S., Donnelly, C. A. & Ferguson, N. M., 2003 Robust parameter estimation techniques for stochastic within-host macroparasite models, *Journal of Theoretical Biology* **225**, 419–430.
- Towers, S., Patterson-Lomba, O. & Castillo-Chavez, C., 2014 Temporal variations in the effective reproduction number of the 2014 West Africa Ebola outbreak, *PLoS Currents Outbreaks* .
- Tuite, A. R., Tien, J., Eisenberg, M., Earn, D. J. D., Ma, J. & Fisman, D. N., 2011 Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial spread of disease and identify optimal control interventions., *Annals of Internal Medicine* **154**, 593–601.
- Wearing, H. J., Rohani, P. & Keeling, M. J., 2005 Appropriate models for the management of infectious diseases, *PLoS Medicine* **2**, e174.
- WHO Ebola Response Team, 2014 Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections, *New England Journal of Medicine* .

Appendix A. Simulation study

To demonstrate the differences between fitting to raw incidence vs. cumulative incidence data, we performed a simulation study in which we fit the deterministic model variant to both types of data at three different levels of observation overdispersion: $k \in \{0, 0.2, 0.5\}$. For each overdispersion treatment, 200 simulated 39-week time series were generated from the stochastic model variant. The basic reproduction number was set to $R_0 = 1.4$; the incubation and infectious periods were fixed as in Table S1; the assumed population size was taken to be that of the Republic of Guinea. We assumed a reporting probability of $\rho = 0.2$ and that, at outbreak initiation, 10 individuals were infected. This set of parameter values yields a sample mean simulation visually comparable to the WHO data from Guinea, which display initially slow growth in the number of cases and later acceleration.

For each simulated data set, we estimated the basic reproduction number, R_0 , the reporting probability, ρ , the negative binomial overdispersion parameter, k , and the initial conditions. All other model parameters were fixed at their true values. Parameter estimation was accomplished using the trajectory matching algorithm (`traj.match`) from the R package `pomp` (King *et al.*, 2010). To maximize the chance of locating the global optimum, the trajectory matching runs were initialized on a latin hypersquare (Sobol’) design of 10^4 points. We used the best fits from this initial run as starting points for additional rounds of optimization to obtain profile likelihoods over R_0 , ρ , and k . From these likelihood profiles, we obtained maximum likelihood point estimates and likelihood-ratio confidence intervals. The full process of obtaining likelihood profiles on model parameters by trajectory matching took approximately 1 cpu hr per simulation.

A second simulation study was performed, in which the deterministic variant of the model was fit to cumulative incidence data by least squares. This common procedure assumes that the measurement error are independent and identically normally distributed. Results of this exercise are shown in Fig. S1 in a form comparable to that of Fig. 1. As in the results shown in the main text, confidence interval widths are erroneously under-estimated with the result that achieved coverage is far smaller than its nominal value.

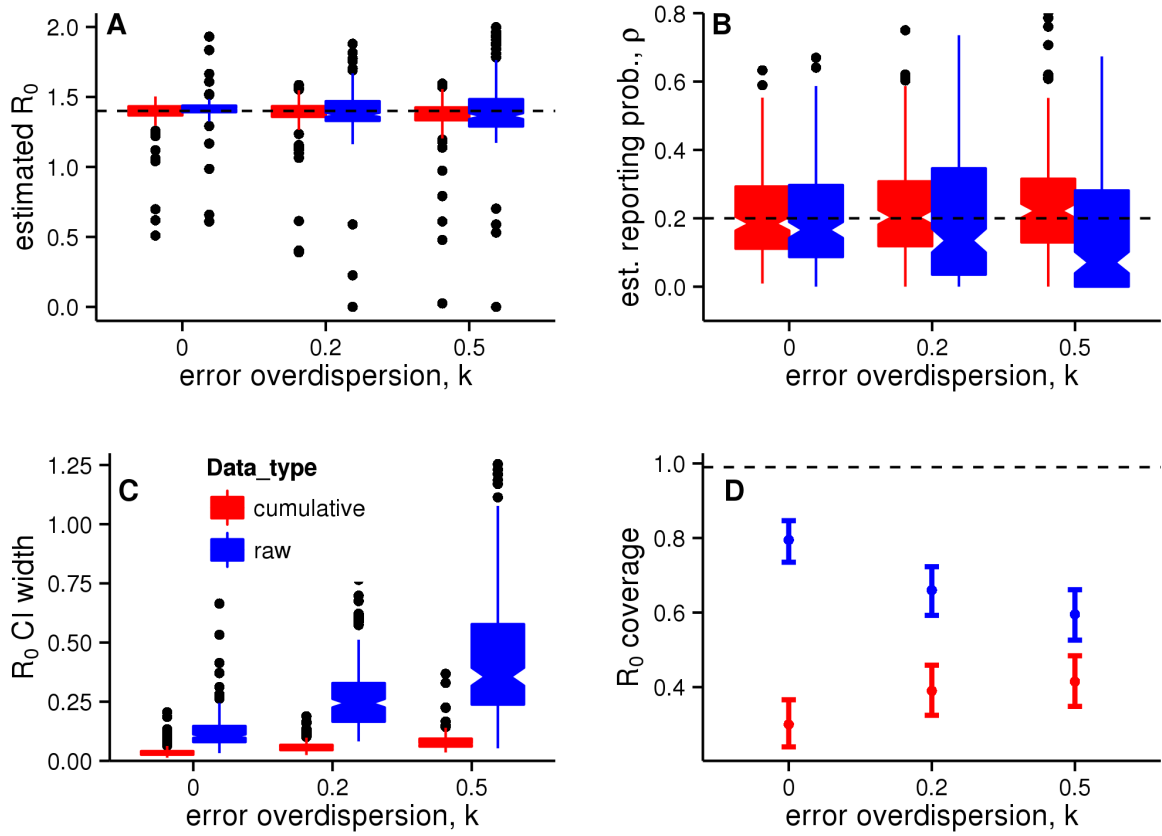


Figure S1: Results from simulation study fitting the deterministic model to cumulative incidence data using the method of least squares. The model was fit to both raw (blue) and accumulated (red) simulated incidence data. The same 200 simulated data sets of length 39 wk used in Fig. 1 were used here. (A) Estimates of R_0 . True value used in generating the data is shown by the dashed line. (B) Estimates of reporting probability, ρ . The dashed line shows the value used to generate the data. (C) Widths of nominal 99% profile likelihood confidence intervals (CI) for R_0 . (D) Actual coverage of the CI, i.e., probability that the true value of R_0 lay within the CI. Ideally, actual coverage would agree with nominal coverage (99%, dashed line).

Appendix B. Model-based inference

Trajectory matching

Model parameters were initially estimated using trajectory matching. As in the simulation study, we fitted R_0 , ρ , k and the initial conditions. All other model parameters were fixed at the known values given in Table S1. The trajectory matching runs were initialized on a Sobol design of 10^5 points. We used the best fits from this initial run to then estimate good starting points in deriving profiles over R_0 , ρ and k . The profile over ρ is relatively flat for $\rho \in (0, 1)$, as it trades off with the initial conditions.

Symbol	Meaning	Value	Citation
R_0	Basic reproduction number	Estimated	
$1/\alpha$	Average incubation period	11.4 da	WHO Ebola Response Team (2014)
m	Incubation period shape parameter	3	WHO Ebola Response Team (2014)
$1/\gamma$	Average infectious period	7 da	WHO Ebola Response Team (2014)
ρ	Reporting probability	Estimated	
k	Reporting overdispersion	Estimated	
N	Population size	Guinea: 10.6M Liberia: 4.1M Sierra Leone: 6.2M	

Table S1: Model parameters, with their interpretations, and their assumed values (parameters estimated from incidence data are so indicated) together with the source of evidence for the assumption.

Iterated filtering

Model parameters were estimated using the Iterated Filtering algorithm (IF) ([Ionides et al., 2011, 2006](#)), implemented as `mif` in the R package `pomp` ([King et al., 2010](#)). For each country, 500 points were generated using a Sobol design and used as starting values for the IF runs (algorithmic parameters: 60 iterations, 2×10^3 particles, hyperbolic cooling with cooling fraction of 0.1, random walk standard deviation, 0.02 for all parameters). For the parameters estimated in each IF run, the log-likelihood was computed as the log of the mean likelihoods of 10 replicate filters, each with 5×10^3 particles. Approximate confidence intervals were then computed using the profile log-likelihood ([Raue et al., 2009](#)). The parameter being profiled was fixed at different values on a predefined grid, and the likelihood maximized over the remaining parameters. To assess Monte-Carlo error in estimation, 20 IF runs were performed from each starting point; the profile log-likelihood was computed as the maximum log-likelihood over these 20 replicates. Computing each of the profile likelihoods in Fig. 2 using iterated filtering took approximately 400 cpu hr of computation, accomplished in roughly 4 hr on a 100 cpu computing cluster.

Parameter	Guinea	Liberia	Sierra Leone	West Africa
R_0	1.22 (1.15–1.29)	1.95 (1.66–2.17)	1.33 (1.21–1.45)	1.48 (1.42–1.54)
ρ	0.01 (0.01–0.07)	0.02 (0.01–0.26)	0.17 (0.11–0.51)	0.004 (0.001–0.015)
k	0.39	0.22	0.07	0.19

Table S2: Parameter estimates. MLE point estimates with nominal 95% confidence intervals are shown.

Additional diagnostics

Fig. S2 shows a comparison of various summary statistics (“probes”) computed both on the data and on model simulations.

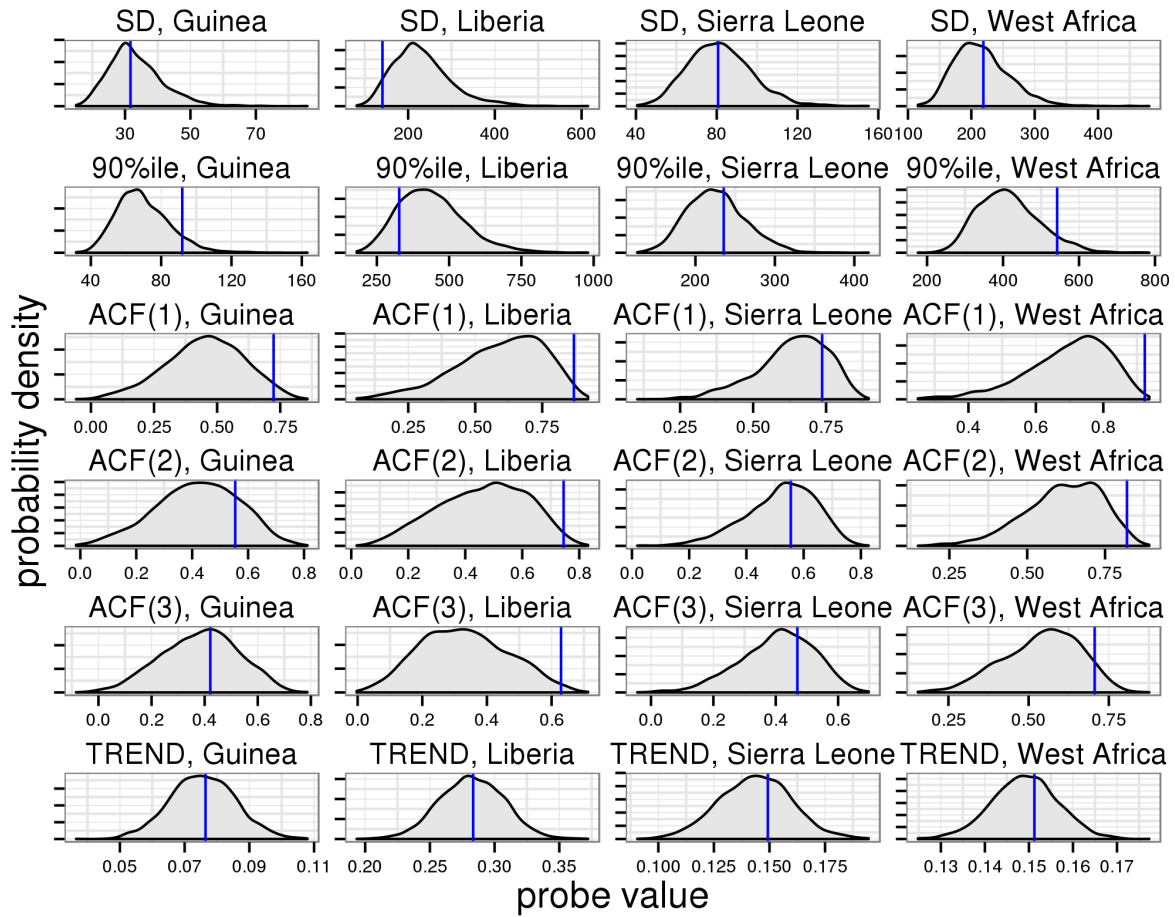


Figure S2: Additional summary statistics, or probes, computed on both stochastic model simulations and the data. Probes include standard deviation (SD), 90th percentile, the autocorrelation at lags 1, 2, and 3 wk (ACF), and the exponential growth rate as obtained by log-linear regression (TREND).