Stochasticity and the limits to confidence when estimating $\mathcal{R}_0$ of Ebola and other emerging infectious diseases - Submission to PLOS Journals

Bradford P. Taylor[1,*], Jonathan Dushoff[2,**], Joshua S. Weitz[3,***]

1 School of Physics, Georgia Institute of Technology, Atlanta, GA, USA
2 Department of Biology and Institute for Infectious Disease Research, McMaster University, Hamilton, Canada
3 School of Biology and School of Physics, Georgia Institute of Technology, Atlanta, GA, USA

*** bradfordptaylor@gmail.com**
****dushoff@mcmaster.ca**
*****jsweitz@gatech.edu**

Abstract

Dynamic models - often deterministic in nature - were used to estimate the basic reproductive number, $\mathcal{R}_0$, of the 2014-5 Ebola virus disease (EVD) epidemic outbreak in West Africa. Estimates of $\mathcal{R}_0$ were then used to project the likelihood for large outbreak sizes, e.g., exceeding hundreds of thousands of cases. Yet fitting deterministic models can lead to over-confidence in the confidence intervals of the fitted $\mathcal{R}_0$, and, in turn, the type and scope of necessary interventions. In this manuscript we propose a hybrid stochastic-deterministic method to estimate $\mathcal{R}_0$ and associated confidence intervals (CIs). The core idea is that stochastic realizations of an underlying deterministic model can be used to evaluate the compatibility of candidate values of $\mathcal{R}_0$ with observed epidemic curves. The compatibility is based on comparing the distribution of expected epidemic growth rates with the observed epidemic growth rate given "process noise", i.e., arising due to stochastic transmission, recovery and death events. By applying our method to reported EVD case counts from Guinea, Liberia and Sierra Leone, we show that prior estimates of $\mathcal{R}_0$ based on deterministic fits appear to be more confident than analysis of stochastic trajectories suggests should be possible. Moving forward, we recommend including a hybrid stochastic-deterministic fitting procedure when quantifying the full $\mathcal{R}_0$ CI at the onset of an epidemic due to multiple sources of noise.

Author Summary

Estimates of the reproductive number, $\mathcal{R}_0$, inform policy-holders about the potential magnitude of an emerging epidemic. These estimates are often derived by leveraging mathematical models where noise in the data and the modeling framework contribute to uncertainty in the estimates. The first estimates of $\mathcal{R}_0$for the 2014-5 Ebola epidemic in West Africa were overly confident partially because process noise, i.e., demographic stochasticity, was not included in the analysis. Process noise is inherent to epidemics due to the discrete nature of disease transmission between individuals. Here we develop a computational methodology for estimating the uncertainty of estimates of $\mathcal{R}_0$due to

process noise alone by leveraging common deterministic fits to simulated stochastic trajectories. We demonstrate how process noise can be a major factor at the onset of an epidemic resulting in a large variance of estimates of epidemiological parameters derived from case data. As a proof of concept, we show that the initial estimates of $\mathcal{R}_0$for the 2014-5 Ebola epidemic were overconfident and, instead, provide a lower bound on the uncertainty of $\mathcal{R}_0$estimates.

Introduction

The SEIRD model of Ebola virus disease (EVD) dynamics, introduced by Legrand and colleagues [1], considers the transitions among Susceptible, Exposed, Infectious, Recovered and Dead (but unburied) individuals. Variants of this core model have been utilized to estimate the basic reproductive number, $\mathcal{R}_0$, of EVD from incidence and cumulative case data in the 2014-5 epidemic in West Africa (e.g., [2–10]). In some instances these estimates include 95% confidence intervals. For example, one of the first mathematical epidemiology papers published in response to the ongoing epidemic in W. Africa estimated the $\mathcal{R}_0$ of EVD to be: 1.51 (95% CI 1.50–1.52) in Guinea; 1.59 (1.57–1.60) in Liberia; and 2.53 (2.41–2.67) in Sierra Leone [2]. The confidence limits appear over-confident, particularly for Guinea and Liberia.

Subsequently, the World Health Organization (WHO) Ebola Response Team analyzed a model with pre- and post-death transmission and estimated $\mathcal{R}_0$ to be: 1.71 (1.44–2.01) for Guinea; 1.83 (1.72–1.94) for Liberia; and 2.02 (1.79–2.26) for Sierra Leone [5]. Similarly, a follow-up model for EVD in the Montserrado region of Liberia estimated $\mathcal{R}_0$ to be 2.49 (2.38–2.60) [8]. Comparison of these case studies shows that models can provide incompatible inferences with non-overlapping CIs despite similar data. Obviously, differences will arise in model structure, data range and quality, and the treatment of noise. Yet, are even these more recent studies over-confident about the precision of $\mathcal{R}_0$ estimates for EVD?

Aaron King and colleagues posed a similar question and cautioned that reported confidence intervals (CIs) of $\mathcal{R}_0$ for EVD are likely too narrow whenever they neglect stochasticity in the disease transmission process [11]. Such over-confidence is further heightened by fitting deterministic models to cumulative case curves (CCCs). A CCC is a monotonically increasing function of time, representing the total number of individuals reported to have become infected during an outbreak. Individual time points within CCC-s are not independent, and so "error" in fitting deterministic models can appear artefactually low. Such low error in fits – if not properly accounted for – can lead to a misleading interpretation of overly narrow CIs for $\mathcal{R}_0$. Instead King et al. [11] recommend a stochastic-based fitting procedure to incidence case curves (ICCs), to account for observation noise and process noise into estimates of $\mathcal{R}_0$ and its associated CIs.

Here, we propose a hybrid stochastic-deterministic approach to address similar issues. The key differentiating feature of our approach from that proposed by King et al. [11], is that we address uncertainty in generation-interval distributions and its effect on $\mathcal{R}_0$. The generation-interval distribution, $g(a)$, is the normalized fraction of secondary cases caused by an infectious individual at "age" a since infection. In the case of EVD, there was – and is – uncertainty with respect to transmission event times, including that of post-death transmission [12, 13].

We apply our hybrid approach retrospectively to estimate the CIs for $\mathcal{R}_0$ for EVD in the summer of 2014, coincident with the release of the first projections for the potential size of the outbreak. We conclude that many early estimates of the CIs for $\mathcal{R}_0$ for EVD were almost certainly over-confident. The over-confidence arose, at least in part, because estimates did not account for the effects of process noise and for uncertainty in

disease generation times. More generally, we explain how models of disease transmission can be adapted to our approach – by using a principled filtering method to identify ensembles of simulated stochastic trajectories "compatible" with a single, measured epidemic outbreak.

Methods

Measuring Compatible Epidemic Trajectories given Stochasticity

Dynamic models provide a means to estimate the severity of an infectious disease from measurements of infectious case data. The severity is usually quantified in terms of the basic reproduction number, $\mathcal{R}_0$– the total number of secondary cases caused by a single infectious individual in an otherwise naive population. An alternative metric of severity is the growth rate of disease incidence, r_0, which is the reciprocal of the characteristic time of exponential growth, τ_c. These quantities are closely related to the doubling time $\tau_2 = \tau_c \ln 2 = \frac{\ln 2}{r_0}$. These two metrics of disease outbreak severity, $\mathcal{R}_0$ and r_0, are related but not equivalent; they are linked by the generation-interval distribution, $g(a)$. For example, in a simple SIR model then $g(a) = \frac{1}{T}e^{-a/T}$ where the average infectious period is T. As is well known, the theoretically expected epidemic growth rate is $r_0 = (\mathcal{R}_0 - 1)/T$. If the generation-interval distribution is known, an estimate of r_0 implies a *unique* corresponding estimate of $\mathcal{R}_0$, e.g., it is $\mathcal{R}_0 = 1 + r_0 T$ in the case of a SIR model. Conversely, when the generation-interval distribution is uncertain, then a range of values of $\mathcal{R}_0$ may be compatible with a given rate of increase in disease incidence – in particular, an increase in the estimated mean generation-interval T would lead to an increase in estimated $\mathcal{R}_0$. This identifiability problem linking the measured value of r_0 and the estimated value of $\mathcal{R}_0$ hampers efforts to estimate the potential scope of the disease outbreak over the long-term [12, 14].

Uncertainty in generation intervals is one of many challenges in $\mathcal{R}_0$. Other important sources of uncertainty include the intrinsic stochasticity of the epidemic, the case-reporting process and uncertainty in the structural mode of disease transmission. Ignoring or under-estimating uncertainty can lead to over-confident estimates of $\mathcal{R}_0$. The premise of our approach is that a measured time series of infectious case data represents a single trajectory from an ensemble of stochastic trajectories given a set of underlying and unknown parameters. The observed trajectory of disease incidence need not grow exponentially at theoretically expected values. Rather there are many trajectories that could appear statistically indistinguishable from the observed epidemic outbreak. We term these trajectories: "**Com**patible **E**pidemic **T**rajectories given **S**tochasticity" or COMETS. The expected variation in the growth rates of COMETS is quantifiable and can be used to bound the confidence of an $\mathcal{R}_0$ estimate from a single stochastic trajectory.

The hybrid stochastic-deterministic approach to estimate the CIs of $\mathcal{R}_0$ from synthetic data relies on an an inverse problem approach to regression-based fits. The approach involves the following steps:

- A range of deterministic models is considered that vary in disease-associated parameters, including $\mathcal{R}_0$.
- For each model and fixed parameter set within the range, we simulate an ensemble of stochastic trajectories and utilize a metric to compare the simulated trajectories to the case data. The metric is the regression-based estimate of the characteristic time, $\hat{\tau}_c$, given either a CCC or an Incidence Case Curve (ICC).

- A value of $\mathcal{R}_0$ is defined to be compatible with a data set if the value of τ_c inferred from the data, $\hat{\tau}_c$, lies in the middle portion of the distribution of values of τ_c inferred from the ensemble of simulations associated with that value, $\tilde{\tau}_c$. Throughout this paper, we focus on the middle 95% of the distribution, and use the symbols $\hat{}$ and $\tilde{}$ to denote estimates from empirical data and simulations, respectively.
- We estimate CIs associated with $\mathcal{R}_0$ by identifying the range of deterministic models that can yield dynamics compatible with the case data, i.e., by identifying COMETS.

This series of steps can be adopted to any epidemic context. Moreover, the well-known problems with inferring CIs for $\mathcal{R}_0$ based on the quality of model fits to a single CCC do not arise via the COMETS approach. In the next section we illustrate why the method is relevant to EVD by highlighting the extent to which process noise drives uncertainty in the distribution of $\tilde{\tau}_c$ of stochastic epidemic trajectories in a SEIRD framework.

Stochastic trajectories of epidemic outbreaks have substantial variation in epidemic growth rates

Stochastic variation in the timing of discrete transmission events can generate variation in estimates of the realized growth rate of an epidemic, r_0. The amount of variation depends on the particular disease model, disease parameters, and the time-scale over which r_0 is estimated. As an example inspired by EVD, we consider disease dynamics based on a SEIRD model:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta_I SI/N - \beta_D SD/N, \quad (1)$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \beta_I SI/N + \beta_D SD/N - E/T_E, \quad (2)$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = E/T_E - I/T_I, \quad (3)$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = (1-f)I/T_I, \quad (4)$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = fI/T_I - D/T_D. \quad (5)$$

We assume that the average latent period is $T_E = 11$ days, the average infectious period is $T_I = 6$ days, there is a $f = .7$ chance of an infected individual ultimately dying and there is an average time of $T_D = 4$ days before burial. We set the disease transmission rates to be $\beta_D = 0.20$ and $\beta_I = 0.25$, meaning that a fraction $\rho_D = \frac{\mathcal{R}_0(\text{dead})}{\mathcal{R}_0} = 0.25$ of transmission is attributable to post-death transmission (see Supplementary Text).In a deterministic framework, epidemics that obey such a model given those parameters should increase with a characteristic time of $\tau_c = 21$ days [12,15].

We simulate an ensemble of 10^4 stochastic realizations of this SEIRD model beginning with a single infectious individual (see Supplementary Text).We consider only realizations that produce at least 50 total cases (approximately 58% of all realizations). This threshold acts as a trigger, from which point we track the time series of incidence, accumulating cases at exact daily intervals. We do not account for reporting error in this example. We estimate the growth rate by fitting an incidence curve (CCC or ICC) from the trigger time t_0 until $t_f = t_0 + 2\tau_c$. For example, in the case of $\tau_c = 21$ then the fitting period is 42 days in duration. The measured epidemic growth rate, $\hat{r}_0$, for a given trajectory is the slope of the best-fit line to log-transformed censused time series assuming errors are Poisson distributed.

Figure 1. Stochastic realizations of a SEIRD epidemic include substantial variability in epidemic growth rate. (Left) Each panel denotes a randomly chosen trajectory for which the cumulative case count exceeded 50 (gray period) followed by a 42 day measurement period (black period). The estimated $\tilde{\tau}_c$ in the measurement period is denoted in the upper left of each panel. (Top Right) Variation in cumulative case counts once a threshold is reached. The results are for 5817 epidemics where the measurement time denotes the period after the first day the trigger was crossed. The solid blue line denotes the median number of cases and the shaded region the central 95% of simulations. (Bottom Right) Variation in the characteristic time estimated from simulation data, $\tilde{\tau}_c$. In all cases, simulations correspond to stochastic simulations of a SEIRD model in which $N = 10^6$, $\beta_I = 0.25$, $\beta_D = 0.2$, $\sigma = 1/11$, $\gamma = 1/6$, $f = 0.7$ and $\rho_D = 0.25$. The theoretically expected characteristic time is $\tau_c = 21$ days. The median characteristic time is $\tilde{\tau}_c = 20.73$ days.

The left panel of Fig. 1 shows five examples of stochastic epidemic trajectories. The time from the index case until the point at which the epidemic "takes off" (i.e., reaches a threshold number of cases), is highly variable. The upper right panel of Fig. 1 shows that, even after take-off, there is substantial variation in growth rate between different epidemic realizations. The trajectories have been time shifted so that $t = 0$ is now defined as the day where the trigger population is reached or surpassed. There is over 40% variation above and below the number of infectious cases at $t = 42$ days. The bottom right panel of Fig. 1 shows the variation in characteristic times estimated from realized data, within an ensemble of epidemic trajectories. The median characteristic time is 20.73 days with a CI of $[17.0, 30.1]$ days. The distribution of the characteristic time widens as the cumulative trigger population decreases (see Fig. S1). This exploratory analysis reveals substantial variability in the growth rate of epidemic trajectories. Quantifying the uncertainty in the characteristic time of epidemic outbreaks in an ensemble is the basis for estimating the CI for $\mathcal{R}_0$.

The expected uncertainty in $\mathcal{R}_0$ for an EVD-like outbreak as estimated from a single stochastic trajectory

Here, we estimate the uncertainty in $\mathcal{R}_0$ for an EVD-like outbreak with SEIRD dynamics. As in the previous section, the fraction of post-death transmission is assumed to be $\rho_D = 0.25$. Then, β_I and β_D are varied to yield different expected characteristic times for the case counts, τ_c (see analytic relationships in Supplementary Text). For each parameter set, we simulate 10^4 stochastic trajectories with the same censusing conditions as in the previous section, conditioned on the fact that the epidemic continues throughout the sampling period. Variation in the measured $\tilde{\tau}_c$ given a range of theoretically expected values of τ_c is shown in the left panel of Fig. 2. This figure is the basis for identifying COMETS and, in turn, for estimating the CIs for $\mathcal{R}_0$ at the onset of the epidemic.

The conventional way to interpret Fig. 2 is as a "forward problem" – as introduced in the previous section. In the forward problem approach, Fig. 2 depicts variation in the measured characteristic time of stochastic epidemic trajectories, $\tilde{\tau}_c$, given variation in the theoretically expected τ_c. The variation is summarized as a probability distribution $p(\tilde{\tau}_c|\tau_c)$. The central 95% of the distribution of $p(\tilde{\tau}_c|\tau_c)$ covers a range whose relative magnitude increases with τ_c.

The alternative way to interpret Fig. 2 is as an inverse problem. In the inverse problem approach, a measurement of any given $\tilde{\tau}_c$ from a single trajectory is compatible with a range of theoretically expected τ_c values. For example, given measurement of $\tilde{\tau}_c = 20$ days, then the associated CI of compatible τ_c is obtained by scanning

Figure 2. Fundamental limits on inferring $\mathcal{R}_0$ from a single stochastic epidemic trajectory. Variation in the CIs of theoretical (left) τ_c and corresponding (right) $\mathcal{R}_0$ when using CCCs. See text for details on methods and interpretation.

horizontally (the red line) for intersections with the forward problem distributions (the black lines). This intersection is estimated by linear interpolation of the 2.5% and 97.5% CI for $p(\tilde{\tau}_c|\tau_c)$ across different values of τ_c. In this example, we estimate the CI of τ_c to be 16.6–27.9 days given a single measurement of $\tilde{\tau}_c = 20$ days.

The CIs for τ_c from the case data is directly translated to CIs for $\mathcal{R}_0$ by utilizing a generating function approach [15]. This method involves determining the moment generating function associated with the distribution of the age of secondary infections resulting from a single infectious individual in an otherwise susceptible population. This distribution varies with τ_c and specifies the initial deterministic dynamics. Converting the τ_c CI to $\mathcal{R}_0$ CI is shown on the right panel of Fig. 2. For the mock case data with $\hat{\tau}_c = 20$ we obtain $\mathcal{R}_0 = 2.10$ with a CI of $1.80 - 2.45$. In summary, these uncertainty ranges arise solely due to process noise and were identified by quantifying COMETS in the SEIRD model framework.

Confidence intervals for $\mathcal{R}_0$ for the EVD outbreak in W. Africa

We now apply the hybrid stochastic-deterministic approach to estimate the confidence limits in $\mathcal{R}_0$ for early outbreak dynamics of EVD in Guinea, Liberia and Sierra Leone. The present method differs from that of the previous section in one key way. Previously, we varied rates of transmission to yield models with different theoretical τ_c. Here, we vary rates of transmission as well as the proportion of post-death transmission, ρ_D. Due to identifiability issues, measured values of $\hat{\tau}_c$ can correspond to different $\mathcal{R}_0$ depending on the underlying parameters of the model [12, 14]. Hence, we systematically vary $\mathcal{R}_0$ between ensembles rather than first varying τ_c and then transforming these into ranges of $\mathcal{R}_0$. We construct the ensembles in a pseudo-Bayesian framework. Specifically, we account for uncertainty in the fraction of infections transmitted the deceased by considering a uniform distribution of values of ρ_D between $0.1 - 0.4$ (see Supplementary Text).In this way, we account for uncertainty in the distribution of times to secondary transmissions in addition to uncertainty from process noise.

Case data was obtained from the WHO [16]. We use cumulative counts and consider dynamics in each country given a start date at which at least 50 cumulative infections have occurred and a final date of September 7, 2014. We choose this time period to reflect the onset of the epidemic across all three countries. The measured characteristic time, $\hat{\tau}_c$ is obtained by fitting an exponential to the CCC for each country after the trigger population is reached. The lowest value of cumulative case counts above 50 for each country is also considered the trigger population for the stochastic simulations. We consider a SEIRD model with a gamma distributed exposed period with $n = 2$ classes in accordance with previous analysis [5]. The ensemble of simulated stochastic trajectories are conditioned on the epidemic remaining throughout the census period, as was the case with the EVD case data. The fits to the stochastic trajectories are subject to the same conditions as the case data.

Resulting CIs of the characteristic time for Liberia, Sierra Leone, and Guinea are shown in the left column of Fig. 3. The characteristic time measured from the case data specifies the location of the red line along the y-axis. For each value of theoretical $\mathcal{R}_0$ we obtain a distribution of measured τ_c from 10^4 simulated stochastic trajectories. The confidence intervals for the reproductive number are determined by where the red line intersects the lower and upper limits of central 95% characteristic time distribution for each theoretical $\mathcal{R}_0$. Note, the conditioning of the epidemic to remain throughout

Figure 3. Estimated $\mathcal{R}_0$ CIs determined by the intersection of the measured case data $\hat{\tau}_c$ with the measured $\hat{\tau}_c$ CIs across a range of theoretical $\mathcal{R}_0$ for each country. (Left column) Characteristic times estimated from case data are given by the height of the red line on the y-axis and the intervals are determined by the intersection of the red line with boundary of the CIs projected onto the x-axis. The blue line refers to the median of the distributions. (Top) Guinea, (Middle) Liberia, (Bottom) Sierra Leone. Table 1 shows the quantitative results for the CIs. (Middle Column) Cumulative case data with projections based on $\mathcal{R}_0$ CIs. The blue line corresponds to the estimate based fitting the case data. The gray area delineates the middle 95% of the projection. The green diamonds are the reported cumulative case data. (Top) Guinea (Middle) Liberia and (Bottom) Sierra Leone. (Right Column) Incident case data with projections based on $\mathcal{R}_0$ CIs. The blue line corresponds to the estimate based fitting the case data. The gray area delineates the middle 95% of the projection. The green diamonds are the reported incident case data. (Top) Guinea (Middle) Liberia and (Bottom) Sierra Leone.

censusing non-negligibly skews the underlying τ_c distributions for low $\mathcal{R}_0$. This is because parameters corresponding to high τ_c lead to stochastic dynamics in which the epidemic ends before the final census point. The dynamics we condition for will more often have larger τ_c by statistical chance. Overall, this gives lower values for the upper bound of the τ_c CIs in those cases. However, this skewing is non-negligible only for models with $\mathcal{R}_0 \approx 1$. Since the upper CI of τ_c corresponds to the upper CI of $\mathcal{R}_0$ our overall CI results will be relatively unaffected. Country-specific estimates of $\mathcal{R}_0$ and associated CIs are shown in Table 1. Again, these CIs represent the combined uncertainty of estimating $\mathcal{R}_0$ from a single stochastic trajectory given uncertainty in the distribution of times to secondary infection. Hence, they represent a lower bound of uncertainty in $\mathcal{R}_0$ given additional uncertainty arising from observation noise. These CIs are larger than many early estimates, e.g., [2].

In the middle (right) column of Fig. 3 we project forward the uncertainty in the cumulative (incident) case counts based on our CIs. We project forward by 8 weeks to show the range of expected case counts over time assuming no further control measures are implemented. The case data lies within the projection for Guinea and Sierra Leone, but are substantially lower for Liberia, likely because transmission parameters had already changed substantially, due to behavior change, control measures, or both, or due to structural differences in the epidemic process [17]. Overall, we expect large uncertainty in projected case counts during the exponential phase of the epidemic, due to process noise and the generation-interval uncertainty alone.

Table 1. Confidence intervals of the SEIRD model given EVD case data in Liberia, Sierra Leone, and Guinea.

Country	$\mathcal{R}_0$	$\mathcal{R}_0$ **CI**
Liberia	2.06	$1.92 - 2.27$
Sierra Leone	1.71	$1.39 - 1.81$
Guinea	1.24	$1.04 - 1.42$

The $\mathcal{R}_0$ values are obtained from fitting a Poisson regression to the case data. The CIs are calculated based on Poisson fits to an ensembles of 10^4 stochastic trajectories. The models used to obtain the trajectories varied in uniformly in ρ_D between [.1, .4].

Discussion

A stochastic implementation of an epidemic leads to significant variation in the realized time series of infectious individuals, due to "process noise" [18] – the stochastic sequence of discrete events in which individuals become infected, infect others, and eventually recover or die. The difference between trajectories predicted in deterministic vs. stochastic models has been observed and studied for decades in other disease contexts [19–22]. Here, we explored the practical consequences of such differences when trying to infer epidemiological parameters, including $\mathcal{R}_0$, at the early stages of an epidemic, and showed that stochastic variation constrains the extent to which confidence limits in $\mathcal{R}_0$ can be narrowed.

In practice, time series that are used for fitting dynamical epidemiological models are drawn from the early stages of an epidemic. At such early stages, a single trajectory may be well fit by a single exponential rate from which $\mathcal{R}_0$ can be estimated. Yet, the trajectories within an ensemble generated given the same underlying parameters will be fit by a distribution of exponential rates. Hence, the 95% CIs for an estimate of $\mathcal{R}_0$ when using a deterministic model are significantly broadened due to process noise.

This general limitation of fitting deterministic models applies to the study of EVD. Multiple groups have proposed model variants of EVD dynamics, fit deterministic models to case data, and then used such fits to estimate $\mathcal{R}_0$, including associated CIs (e.g., [2, 5, 7, 8]). The time range we used for fitting includes the period from onset as defined by the time with greater than 50 cumulative infections to early September 2014. This range spans from the onset of the epidemic to reported cumulative case counts of nearly 700 in Guinea and over 1500 in both Sierra Leone and Liberia. Using a stochastic implementation of a SEIRD model, we inferred $\mathcal{R}_0$ to be 1.24 for Guinea (1.04-1.42 95% CI), 2.06 for Liberia (1.95-2.29 95% CI), and 1.71 for Sierra Leone (1.4-1.81 95% CI). The estimates and ranges reflect three assumptions: (i) the model structure; (ii) the prior assumptions about the exposed and infectious periods, time to bury and probability of death; (iii) the availability and quality of epidemic case data. In particular, the second assumption involves varying the distribution of time to secondary infection. These CIs denote limits to inference in this class of model imposed by the nature of the data available: a single stochastic epidemic outbreak. The original estimates of CIs from [2] are evidently too narrow, yet even later estimates should be revisited using the hybrid approach proposed here.

The present method is complementary to alternative, profile-likelihood based approaches [11]. We obtain distributions for $\mathcal{R}_0$ by fitting to ensembles obtained by stochastically simulating a range of deterministic models. We marginalized our ensembles so that the trajectories in the ensembles were obtained from underlying models with the same theoretical τ_c. Previous work used particle filtering techniques to obtain a distribution based on fitting to the data [11]. This approach estimates uncertainty due to fitting deterministic models to a single, stochastic trajectory. In our case, we find that differences between fitting to cumulative or incident data are negligible, so long as the quality of individual fits within an ensemble is not used as the primary source of information on the CIs (see Supplementary Text).We note that our methodology of estimating COMETS is similar to the method of "plausible parameter sets" advocated for use in estimating disease-associated parameters during outbreaks [23].

The hybrid approach comes with certain precautions. Nonlinear models often display sloppiness such that many combinations of parameters have little effect on certain system dynamics [24]. The identifiability problem [12, 14] also applies here (see Supplementary Text), so that the strength of fit does not necessarily exclude a range of compatible mechanisms. Implementing the current, hybrid approach should be straightforward to adapt to both well-mixed and spatially-explicit models of disease

transmission. For example, a recent analysis of spatial data suggested that apparent exponential dynamics is a result of aggregation of local epidemics [25]. Yet, we expect that high-performance implementations of stochastic models will be required for spatially explicit projections of outbreak sizes and associated CIs for $\mathcal{R}_0$ at the early stages of an epidemic [26].

In summary, we should have expected to have less confidence in estimates of CIs of $\mathcal{R}_0$ at the outset of a EVD epidemic given process noise. We hope that the current method, similar in intent to that of King et al. [11], provides an accessible route for estimating realistic CIs for $\mathcal{R}_0$ in epidemics. In practice, the 95% confidence intervals in $\mathcal{R}_0$ estimated from stochastic model fits will be broader than that estimated from deterministic model fits to cumulative case data. Estimates of CIs using the current hybrid approach represent a lower bound of uncertainty due to stochastic sources of noise. As an epidemic continues and the number of infected individuals increases, observation noise contributes a relatively larger proportion of uncertainty as compared to process noise [22]. Remaining realistic about the limits to confidence in model fits should also be incorporated into public health practice, e.g., when projecting the necessary scope of intervention based on "optimal" fits [9]. We encourage the academic, governmental, and non-governmental public health communities to consider incorporating unavoidable uncertainty into their decision making pipelines when responding to emergent disease outbreaks.

Supporting Information

S1 Text

Supplementary Text Includes: generating function approach to link characteristic times and transmission rates, methods of stochastic simulations of epidemic outbreaks, pseudo-Bayesian approach for uncertainty in ρ_D, comparing results from CCC and ICC data, issues with identifiability.

S1 Fig

The distribution of the characteristic time, τ_c increases with decreasing trigger population. Ensembles of SEIRD stochastic dynamics are simulated with parameters as in Fig. 2. The ensemble includes 10^4 trajectories with an epidemic that persists for 300 days. In each case, estimates of τ_c are based on fits to the CCC for 42 days after the trigger population is reached.

S2 Fig

Variation in regression-based fits of the characteristic time of case counts between estimates using cumulative or incidence case count data. The theoretical characteristic time of the underlying model is $\tau_c = 20$ as shown by the black dashed vertical line. The ICC distribution has a median of 20.3 days with 95% CIs of 16.5-28.3. The CCC distribution has a median of 19.9 days with 95% CIs of 16.5-28.0.

S3 Fig

Distributions of standard error of estimated growth rates, $r_0 = \hat{\tau}_c^{-1}$ for CCCs and ICCs arising from linearly fitting simulated data assuming deviations are Poisson distributed. The underlying deterministic model for simulated data has a theoretical characteristic time of $\tau_c = 20$ days.

S4 Fig

Identifiability problem persists when fitting SEIRD models to stochastic data. The three scenarios correspond to cases where the characteristic time, $1/r_0$=14, 21 and 28 days. The realized epidemic growth rates of stochastic trajectories are measured given variation in ρ_D from 0 to 1 in increments of 0.1. Circles denote the median characteristic time while triangles denote the 95% confidence intervals from an ensemble of 10^3 simulations per condition.

Acknowledgments

The work was funded by a grant to JSW from the Burroughs Wellcome Fund and from the Army Research Office grant #W911NF-14-1-0402. We would like to acknowledge Luis F. Jover for reviewing the manuscript.

References

1. Legrand J, Grais RF, Boelle PY, Valleron AJ, Flahault A. Understanding the dynamics of Ebola epidemics. Epidemiology and infection. 2007;135(04):610–621.
2. Althaus CL. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. PLoS Currents Outbreaks. 2014;6.
3. Gomes MF, Pastore y Piontti A, Rossi L, Chao D, Longini I, Halloran ME, et al. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. PLoS Currents Outbreaks. 2014;6.
4. Fisman D, Khoo E, Tuite A. Early epidemic dynamics of the West African 2014 Ebola outbreak: estimates derived with a simple two-parameter model. PLoS Currents Outbreaks. 2014;6.
5. WHO Ebola Response Team. Ebola virus disease in West Africa – the first 9 months of the epidemic and forward projections. New England Journal of Medicine. 2014;371(16):1481–1495.
6. Nishiura H, Chowell G. Early transmission dynamics of Ebola virus disease (EVD), West Africa, March to August 2014. Euro Surveill. 2014;19(36):20894.
7. Pandey A, Atkins KE, Medlock J, Wenzel N, Townsend JP, Childs JE, et al. Strategies for containing Ebola in west Africa. Science. 2014;346(6212):991–995.
8. Lewnard JA, Mbah MLN, Alfaro-Murillo JA, Altice FL, Bawo L, Nyenswah TG, et al. Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis. The Lancet Infectious Diseases. 2014;14(12):1189–1195.
9. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, et al. Estimating the future number of cases in the Ebola epidemic-Liberia and Sierra Leone, 2014–2015. MMWR Surveill Summ. 2014;63(suppl 3):1–14.
10. Rivers CM, Lofgren ET, Marathe M, Eubank S, Lewis BL. Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia. PLoS Currents Outbreaks. 2014;6.

11. King AA, Domenech de Cellès M, Magpantay FMG, Rohani P. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. Proceedings of the Royal Society of London B: Biological Sciences. 2015;282(1806).

12. Weitz JS, Dushoff J. Modeling Post-death Transmission of Ebola: Challenges for Inference and Opportunities for Control. Scientific reports. 2015;5.

13. Nielsen CF, Kidd S, Sillah A, Davis E, Mermin J, Kilmarx PH. Improving burial practices and cemetery management during an ebola virus disease epidemic-Sierra Leone, 2014. MMWR Surveill Summ. 2015;64:1–8.

14. Eisenberg MC, Eisenberg JNS, D'Silve JP, Wells EV, Cherng S, Kao YH, et al. Forecasting and Uncertainty in Modeling the 2014-2015 Ebola Epidemic in West Africa. http://arxivorg/pdf/150105555v3pdf. 2015;.

15. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings of the Royal Society B: Biological Sciences. 2007;274(1609):599–604.

16. WHO. WHO, editor. http://apps.who.int/gho/data/node.ebola-sitrep; 2014. [Online; accessed 17-December-2014]. Available from: `http://apps.who.int/gho/data/node.ebola-sitrep`.

17. Chowell G, Simonsen L, Viboud C, Kuang Y. Is West Africa approaching a catastrophic phase or is the 2014 Ebola epidemic slowing down? Different models yield different answers for Liberia. PLoS Currents Outbreaks. 2014;6.

18. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Princeton, NJ: Princeton University Press; 2007.

19. Gibson GJ, Renshaw E. Estimating parameters in stochastic compartmental models using Markov chain methods. Mathematical Medicine and Biology. 1998;15(1):19–40.

20. Ionides EL, Breting AA. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences. 2006;103(49):18438–18443.

21. Cauchemez S, Ferguson NM. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. Journal of The Royal Society Interface. 2008;5(25):885–897.

22. Ma J, Dushoff J, Bolker BM, Earn DJD. Estimating initial epidemic growth rates. Bull Math Biol. 2014;76:245–260.

23. Drake JM, Kaul R, Alexander LW, O'Regan SM, Kramer AM, Pulliam JT, et al. Ebola cases and health system demand in Liberia. PLoS Biol. 2015;13(1):e1002056.

24. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS computational biology. 2007;3(10):e189.

25. Chowell G, Nishiura H. Transmission dynamics and control of Ebola virus disease (EVD): a review. BMC Medicine. 2014;12:196.

26. Merler S, Ajelli M, Fumanelli L, Gomes MF, y Piontti AP, Rossi L, et al. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. The Lancet Infectious Diseases. 2015;15(2):204–211.