Title: The IDEA model: A single equation approach to the ebola forecasting challenge

Author: Ashleigh R. Tuite David N. Fisman

PII: S1755-4365(16)30030-5
DOI: http://dx.doi.org/doi:10.1016/j.epidem.2016.09.001
Reference: EPIDEM 219

To appear in:

Received date: 20-6-2016
Revised date: 23-9-2016
Accepted date: 25-9-2016

Please cite this article as: Tuite, Ashleigh R., Fisman, David N., The IDEA model: A single equation approach to the ebola forecasting challenge.Epidemics http://dx.doi.org/10.1016/j.epidem.2016.09.001

The IDEA Model: A Single Equation Approach to the Ebola Forecasting Challenge

Ashleigh R. Tuite[1] and David N. Fisman[2]

[1] Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Corresponding author:
Ashleigh Tuite
Room 106, 1633 Tremont Street
Boston, MA, USA, 02120
Email: atuite@hsph.harvard.edu
Phone: 617-432-6761

Abstract

Mathematical modeling is increasingly accepted as a tool that can inform disease control policy in the face of emerging infectious diseases, such as the 2014-2015 West African Ebola epidemic, but little is known about the relative performance of alternate forecasting approaches. The RAPIDD Ebola Forecasting Challenge (REFC) tested the ability of eight mathematical models to generate useful forecasts in the face of simulated Ebola outbreaks. We used a simple, phenomenological single-equation model (the "IDEA" model), which relies only on case counts, in the REFC. Model fits were performed using a maximum likelihood approach. We found that the model performed reasonably well relative to other more complex approaches, with performance metrics ranked on average 4^{th} or 5^{th} among participating models. IDEA appeared better suited to long- than short-term forecasts, and could be fit using nothing but reported case counts. Several limitations were identified, including difficulty in identifying epidemic peak (even retrospectively), unrealistically precise confidence intervals, and difficulty interpolating daily case counts when using a model scaled to epidemic generation time. More realistic confidence intervals were generated when case counts were assumed to follow a negative binomial, rather than Poisson, distribution. Nonetheless, IDEA represents a simple phenomenological model, easily implemented in widely available software packages that could be used by frontline public health personnel to generate forecasts with accuracy that approximates that which is achieved using more complex methodologies.

Keywords: Ebola virus disease, mathematical modeling, forecasting

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Highlights

- The “incidence decay with exponential adjustment” (IDEA) model is a single equation, phenomenological model that was used in the RAPIDD Ebola Forecasting Challenge.
- Across a variety of performance metrics model forecasts were comparable to other, more complex and mechanistically explicit models.
- Key challenges identified included accurate forecasting of simulated epidemic peaks, overly precise confidence intervals, and scaling of the model in generation time rather than clock time.
- The optimal approach to use is still evolving, but the IDEA model appears to represent a simple and easily implemented forecasting method that may have application in frontline public health settings.

Introduction

Early in an infectious disease outbreak, data may be limited, and there may be great uncertainty associated with any data that are available. In the case of an emerging infectious disease, whether completely novel or a disease emerging in a new region, there is the further challenge of not having the surveillance systems in place to identify baseline disease activity or (initially) to accurately measure outbreak growth in the population. The West African Ebola epidemic is an example of the devastating toll that delays in identifying and controlling emergent infectious diseases can have on the population.

Once the Ebola outbreak was recognized, understanding the scale and scope of the problem was a challenge due to limited and unreliable data. Nonetheless, decisions had to be made, and clinicians, public health practitioners, and policy-makers made use of the best available data to both try to understand the current situation and project the likely future course of the epidemic. Mathematical modeling was used to generate many of these projections, and the message was stark: in the absence of more effective intervention, we might expect hundreds of thousands of cases Ebola (1-3). Fortunately, the scale of the epidemic never reached the level of those early projections. The precise mechanisms driving the epidemic decline are still debated and may be unknowable (4, 5). Some subsequently criticized mathematical models for generating alarming projections via worst case scenarios, or failure to explicitly model interventions which were not yet extant (6).

The RAPIDD Ebola Forecasting Challenge (REFC) (described in another paper in this issue) provided mathematical modelers with an opportunity to test models of varying complexity for their ability to forecast several simulated Ebola outbreaks. Because these simulated data have an underlying 'truth', and presented scenarios with varying quantities and qualities of information, participants were able to determine if the difficulties associated with projecting the impact of interventions in the actual Ebola epidemic were due to inadequate data or an underlying problem with how mathematical models capture interventions. Our contribution to this exercise involved use of a simple, single equation model that relied solely on national case report data, to generate projections of the simulated outbreaks.

Methods

Modeling Approach

We used the previously described incidence decay with exponential adjustment (IDEA) model (7). Briefly, this single-equation model describes epidemic processes in terms of both first order exponential growth (a function of the basic reproductive number, R_0) and simultaneous second order decay. We are agnostic about the nature of factors that slow growth, but they could be postulated to include behavioural change, public health interventions, increased immunity in the population, or any other dynamic change that slows disease transmission (7). The model is purely descriptive (or "phenomenological") and cannot distinguish between putative controlling mechanisms, but has the advantage of allowing epidemic growth to slow even before the critical fraction of susceptible individuals in the population is exhausted.

The model uses the functional form:

$$I_t = \left(\frac{R_0}{(1+d)^t}\right)^t$$

where t is scaled in generation time (time from infection of an initial to case to infection of his or her secondary case (8)), and d represents a control parameter that causes incidence to decay. I_t represents incident cases in a given generation. In the absence of control, incident case counts grow to the power

of *t*. However, when control is present, the effective reproduction number is reduced by a power of t^2, causing transmission to slow and stop even when the absolute value of *d* is small. The IDEA model can be readily parameterized by fitting to either incidence or cumulative incidence data alone (though the use of the latter has been criticized due to non-independence of observations (9)) or both simultaneously, requires no assumptions regarding immune status in the population, and has only three parameters.

We previously used the IDEA model to describe the early epidemic dynamics of the West African Ebola epidemic and demonstrated the importance of control efforts and vaccine timing (3, 10).

Application to the RAPIDD Ebola Forecasting Challenge

For the REFC, we assumed that the generation time of Ebola was known, and fixed at 15 days (3). For each of the four scenarios, we used the provided national-level weekly case report data. We fit to both cumulative cases (total number of cases occurring up to a given outbreak generation) and incident cases (number of cases occurring at each outbreak generation). Due to the slight dissociation between the 14-day generation time that would be obtained by summing weekly counts and the 15-day generation time we had decided to use *a priori*, cumulative cases were calculated by summing the weekly reported cases and determining the total number of cases that would have been reported at the

end of each epidemic generation. For the same reason, incident cases per generation were calculated by subtracting cumulative cases at the previous generation from the cumulative cases at the current generation. Although additional data were provided for the scenarios, we restricted ourselves to using the national case report data, taking an agnostic approach for the purposes of parameter estimation and providing the required projections.

Although control is implicit in the model formulation, the mechanisms of control are not represented explicitly. As such, we did not attempt to incorporate the information provided in the scenario situation reports to evaluate the contributions of different interventions to the outbreak dynamics.

For each scenario, we estimated the generation of initial recognition of cases based on the number of cases reported at the first weekly data point, assuming an R_0 of approximately 2-3. For scenarios 1-3, we estimated that Ebola had been circulating in the population for two generations before it was reported (i.e., the first reported cases represented the third generation). For scenario 4, we assumed that the initial case was reported.

For each round of the challenge, we estimated the best-fit parameter values for R_0 and d using maximum likelihood methods, using the mle2 function in the bblme R package (11). We assumed observations to be Poisson distributed. The 25^{th} and 75^{th} percentile estimates of R_0 and d (calculated using the confint

function of bbmle) were used to generate bounds for the model projections. Because R_0 and d estimates are positively correlated (i.e., well-fitting models with higher R_0 have higher d) the 75th percentile projections were based on lower bound values for both parameters, while the 25th percentile projections are based on upper bound parameter values for both parameters (10).

The model outputs were rescaled from generation time to weekly estimates by dividing the total number of cases occurring within each generation by the generation time to get daily case counts, and aggregating up to weekly case counts. Peak week was determined as the week with the maximum number of cases or the midpoint if two weeks had the same projected number of cases. Outbreak final size was calculated as the total number of cases occurring from the start of the outbreak until the weekly case count fell below 1.

Performance Metrics

Incidence forecasts were evaluated using six metrics: R^2 and Pearson's correlation coefficient for observed and expected incidence; mean squared error and root-mean squared error; mean absolute error; and mean absolute percentage error. For each scenario, metrics were averaged across all prediction time points and forecast lead times. Additionally, overall summary measures of performance were obtained by averaging metrics across all scenarios. As these metrics are difficult to compare directly, we calculated the rank of the IDEA model

for each metric, for each scenario, as well as overall performance for all 4 scenarios. We then calculated mean rank and range for IDEA relative to other modeling approaches.

Modifications to the Modeling Approach

After the completion of the challenge, we evaluated alternate assumptions about the distribution of underlying data, to better characterize the uncertainty associated with our model projections. Specifically, we considered the impact of assuming a negative binomial distribution rather than a Poisson distribution, as case counts for EVD are reported to be overdispersed (12-15). We also compared projections generated by fitting to incident cases only, as opposed fitting simultaneously to incident and cumulative cases.

Results

An example of how the model fit to the available data, and how the estimated parameter values changed over time is provided in **Supplementary Figure 1**. For each of the four scenarios and estimation time points, we visually compared the model projections of the outbreak trajectory over the subsequent four weeks to what actually unfolded (**Figure 1**). The ability to predict the outbreak trajectory varied from scenario to scenario. For all scenarios, the interquartile range of projected EVD case counts was very small and tended not to overlap with the true case counts.

For each scenario, the timing of the epidemic peak was estimated at various time points throughout the epidemic. For all scenarios except scenario 4, estimates made later in the course of the epidemic more closely approximated the true epidemic peak week (**Figure 2**). At the last estimation time point under scenario 4, our model predicted a downturn in disease transmission that did not occur. As a result, the model incorrectly suggested that the peak had already passed, when in fact it had not yet occurred. As the outbreak progressed, the model provided reasonably accurate estimates of outbreak final size (**Figure 3**), even when, as in scenario 4, it struggled to forecast short-term dynamics.

When we ranked performance for each of six metrics, and for each scenario, we found that IDEA was consistently ranked near the midpoint among the eight models used in the challenge. The best mean rank across performance metrics

was seen in Scenario 2 (mean rank 3.8) and the worst in Scenario 1 (mean rank 4.7); IDEA's mean rank for overall fit was 4.8 (**Supplementary Figure 2**).

When applied to transmission data, a Poisson distribution assumes no variability in contact rates or transmission probability per contact (13, 15). This property gave our model forecasts a degree of precision that did not reflect the underlying outbreak dynamics. Given the apparent importance of superspreading events and individual variability in EVD transmission we refit the model to the data from scenario 1, assuming a negative binomial distribution of cases and using a literature-derived estimate for the dispersion parameter of 0.2 (12) (**Figure 4**). As expected, compared to the Poisson model, this approach resulted in less precise estimates of the future trajectory of the outbreak, generating an interquartile range that overlapped with the actual outbreak trajectory at all time points. Similarly, the interquartile range for the week of the epidemic peak included the actual peak for all time points (data not shown). Our estimates of outbreak final size overlapped with the actual value for all but time point 3, but the uncertainty in these estimates were very large at the early prediction time points (e.g., at time point 1, interquartile range: 700-3.7×10^{10} cases).

Finally, we evaluated the impact of fitting to incident cases only. Compared to the approach we took in our main analysis of fitting to both incident and cumulative case counts, we did not observe a meaningful difference in short-term model projections (**Figure 5**), or in estimated peak timing or final outbreak size.

Discussion

We sought to apply a simple, single equation phenomenological model to the forecasting of simulated Ebola case data. Overall, the performance of our model was comparable to that of other more complex approaches, with the added advantage of simplicity and few assumptions about the mechanisms underlying changing disease dynamics. Indeed, we did not make use of the situation reports provided, nor did we attempt to explicitly model the effects of interventions; our "agnostic" approach may be attractive in the face of real emerging infectious disease outbreaks, since we are able to generate reasonable projections without detailed situational information, and using data of suboptimal quality.

Despite extreme simplicity and our agnostic approach to scenario evaluation, the IDEA model performed in a manner comparable to other more complex and mechanistically explicit models, with average ranks in performance ranging from 4-5 among the eight groups participating in the challenge. We found this encouraging, as the simplicity of our model (it can be recreated in widely available spreadsheet software), combined with reasonable performance, may make it a useful tool for "quick and dirty" forecasting by frontline public health personnel.

For all that our model performed reasonably well, we note several important limitations to the approach we applied. Our approach struggled to identify epidemic peaks, even retrospectively. Accurate identification of peaks is

epidemiologically important, as the peak represents the point subsequent to which effective reproductive numbers are < 1, such that the epidemic will end if these dynamics are maintained, and interventions such as vaccination may have limited impact on herd dynamics (16). While we plan to investigate this limitation further we postulate that it may arise from the fact that the IDEA equation describes a symmetrical curve, but the epidemic curves we were tasked with projecting were, in fact, asymmetrical. In IDEA, R_0 is taken to be a fixed property of the disease in question, but we have previously noted sharp shifts in estimates of *d*, the "control parameter", some of which may reflect abrupt improvement in control measures or in population behaviors. For example, we have noted that it is challenging to capture the abrupt decline in transmission that occurred around October 2014 in the actual West African Ebola epidemic with a single *d* parameter ((17) and unpublished observations). The use of multiple time-specific control parameter values for improved epidemic forecasting is an area of active research for our group.

For the purposes of the challenge, we opted to fit simultaneously to both incident and cumulative case counts, reflecting the fact that during the West African EVD epidemic, cumulative case curves were the primary source of publicly-available data. However, we note that there has been much discussion of the appropriateness of cumulative case curves for model fitting and best-practices would support the use of incident case data for model fitting (9). Although there are theoretical reasons to avoid the use of cumulative incidence data, our

supplementary analysis found little difference in model performance with incident cases alone, suggesting that future applications of this model in outbreak situations should use incident case data, or “pseudo-incidence” data derived by differencing cumulative incidence curves, instead.

A further limitation of our model is that it is scaled in “generation time” rather than continuous or even discrete daily time. With a disease such as Ebola virus infection, when generations are thought to be long (approximately 2 weeks) this makes short-term projections challenging and results in forecasts that are staggered and “step-like”. We are currently evaluating the use of smoothing functions to distribute cases, by day, over the course of a given generation.

Lastly, our assumption that case counts would be Poisson distributed resulted in excessively narrow confidence intervals during the Ebola challenge. We have shown that altering our assumption around the case generation process through the use of a negative binomial distribution in our likelihood function may be more appropriate for describing Ebola and other emerging infections of public health importance that have overdispersed case distributions due to superspreading events (12).

In summary, the IDEA model is a simple, single-equation phenomenological model that performed reasonably well relative to more complex models in the context of the Ebola challenge. Our approach joins logistic models, including the

Richards model, as simple, phenomenological approaches to epidemic modeling and forecasting in real time (18-20), with the added advantage of the highly intuitive nature of the parameters in our model. The model has several limitations, but may nonetheless find application as a simple forecasting tool which complements more complex mechanistic approaches, and which can be easily recreated (e.g., in spreadsheet software) and rapidly applied in front line public health settings, including those with limited computing resources.

Acknowledgements

The authors thank the challenge organizers for their support during the challenge and for providing the performance metrics included in this analysis. The authors are also grateful to the other challenge participants, particularly David Champredon, for feedback and suggestions for improving the model.

References

1. Lewnard JA, Ndeffo Mbah ML, 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. Lancet Infect Dis. 2014;14(12):1189-95.
2. Althaus CL**.** Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa. PLoS Curr. 2014;6.
3. 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 Current Outbreaks. 2014(1).
4. Atkins KE, Pandey A, Wenzel NS, Skrip L, Yamin D, Nyenswah TG, et al. Retrospective Analysis of the 2014-2015 Ebola Epidemic in Liberia. Am J Trop Med Hyg. 2016;94(4):833-9.
5. Kucharski AJ, Camacho A, Flasche S, Glover RE, Edmunds WJ, Funk S**.** Measuring the impact of Ebola control measures in Sierra Leone. Proc Natl Acad Sci U S A. 2015;112(46):14366-71.
6. Butler D**.** Models overestimate Ebola cases. Nature. 2014;515(7525):18.
7. Fisman DN, Hauck TS, Tuite AR, Greer AL**.** An IDEA for short term outbreak projection: nearcasting using the basic reproduction number. PloS one. 2013;8(12):e83622.
8. Svensson A**.** A note on generation times in epidemic models. Math Biosci. 2007;208(1):300-11.
9. King AA, Domenech de Celles M, Magpantay FM, Rohani P**.** Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. Proc Biol Sci. 2015;282(1806):20150347.
10. Fisman D, Tuite A**.** Projected impact of vaccination timing and dose availability on the course of the 2014 west african ebola epidemic. PLoS Curr. 2014;6.
11. Bolker B**.** bblme: tools for general maximum likelihood estimation. Available at: https://cran.r-project.org/web/packages/bbmle/index.html. 2016.
12. Althaus CL**.** Ebola superspreading. Lancet Infect Dis. 2015;15(5):507-8.
13. Toth DJ, Gundlapalli AV, Khader K, Pettey WB, Rubin MA, Adler FR, et al. Estimates of Outbreak Risk from New Introductions of Ebola with Immediate and Delayed Transmission Control. Emerg Infect Dis. 2015;21(8):1402-8.
14. Faye O, Boelle PY, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis. 2015;15(3):320-6.
15. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM**.** Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438(7066):355-9.

16. Pandemic Influenza Outbreak Research Modelling T, Fisman D**.** Modelling an influenza pandemic: A guide for the perplexed. CMAJ. 2009;181(3-4):171-3.
17. Nyenswah TG, Westercamp M, Kamali AA, Qin J, Zielinski-Gutierrez E, Amegashie F, et al. Evidence for declining numbers of Ebola cases--Montserrado County, Liberia, June-October 2014. MMWR Morb Mortal Wkly Rep. 2014;63(46):1072-6.
18. Chowell G, Hincapie-Palacio D, Ospina J, Pell B, Tariq A, Dahal S, et al. Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics. PLoS Curr. 2016;8.
19. Hsieh YH, Ma S**.** Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005. Am J Trop Med Hyg. 2009;80(1):66-71.
20. Xiao Y, Patel Z, Fiddler A, Yuan L, Delvin ME, Fisman DN**.** Estimated impact of aggressive empirical antiviral treatment in containing an outbreak of pandemic influenza H1N1 in an isolated First Nations community. Influenza Other Respir Viruses. 2013;7(6):1409-15.

Figure Legends

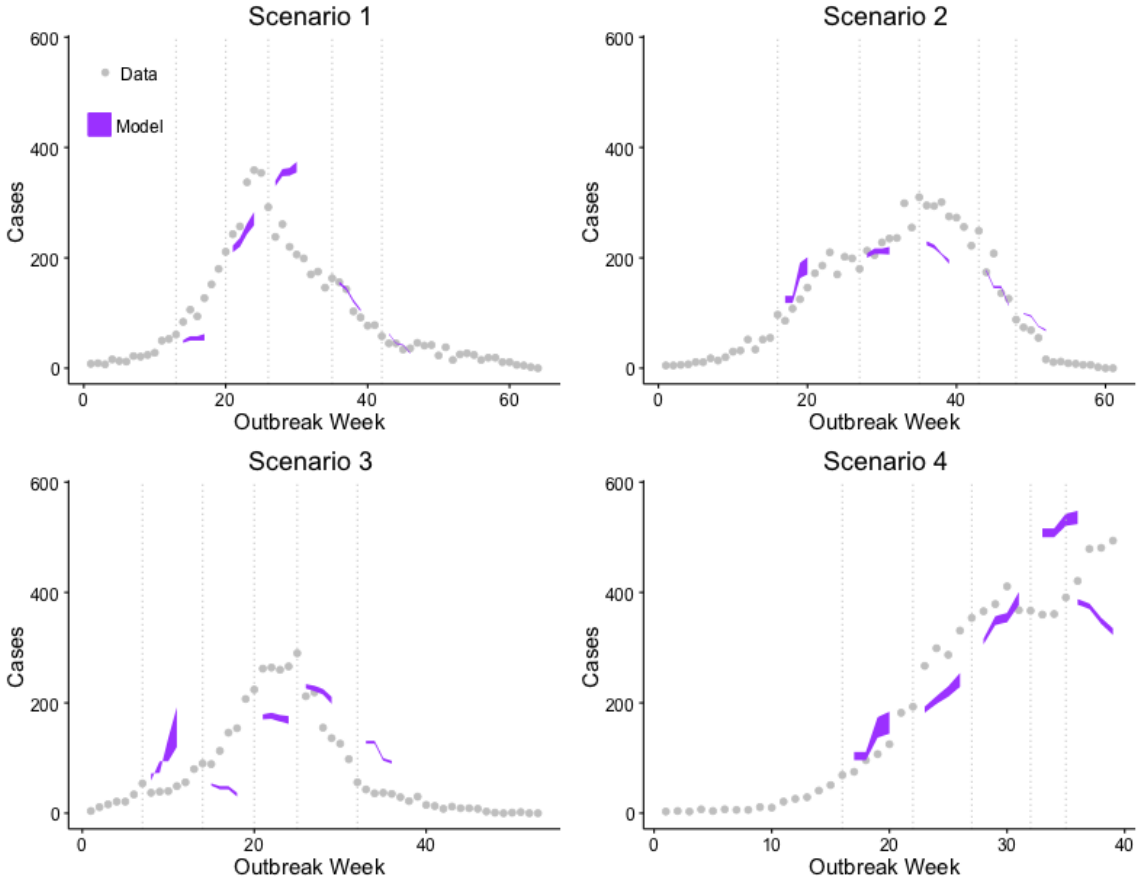

Figure 1. Model projections compared to simulated national outbreak data. Model projections were generated at 5 time points (indicated by dashed vertical lines), using the simulated outbreak data (black circles) available up until to the prediction week. The number of cases was projected for up to 4 weeks after each time point, with the interquartile range of these projections shown in purple.

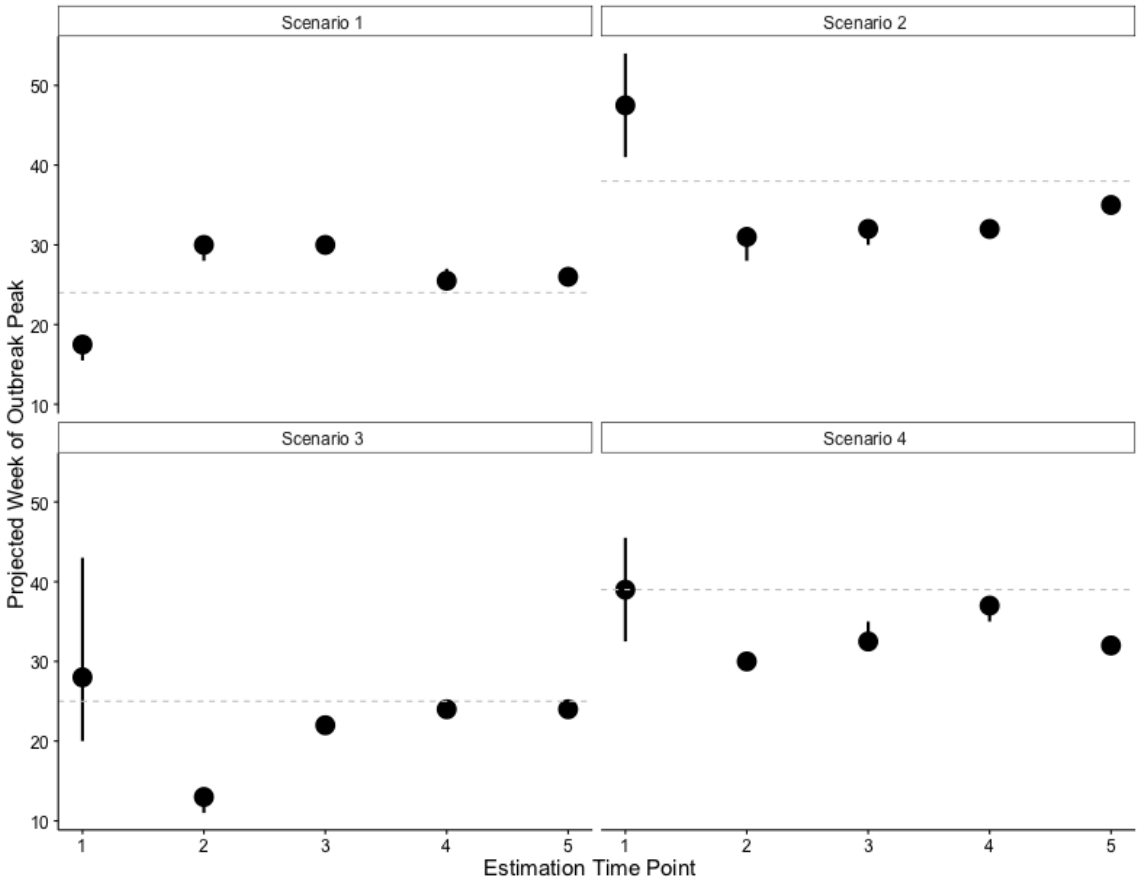

Figure 2. Model-projected timing of the epidemic peak. The timing of the peak was estimated at each of the 5 time points, for the 4 scenarios. Circles represent the mean estimated peak, with error bars showing the interquartile range. Dashed horizontal lines show the actual timing of the outbreak peak for each scenario.

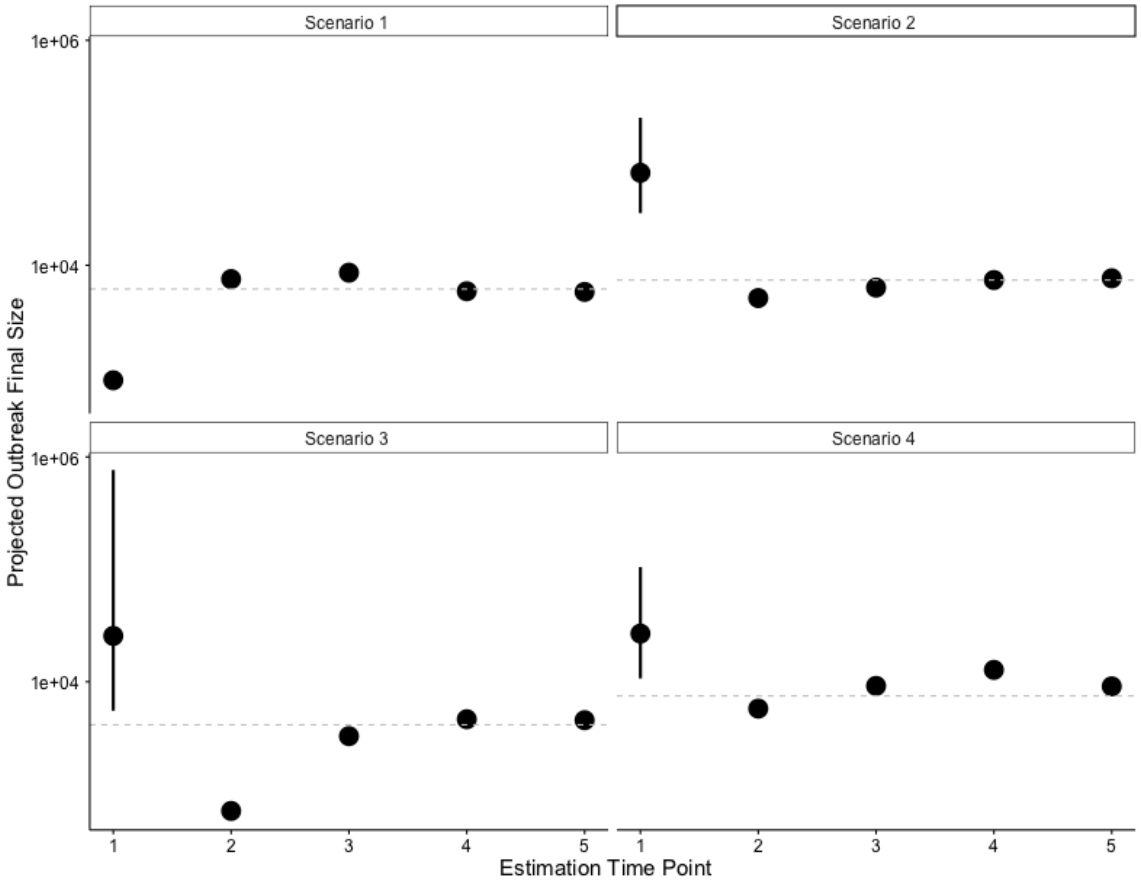

Figure 3. Model-projected outbreak final size. The expected final size of the outbreak was estimated at each estimation time point for the 4 scenarios. Circles indicate the mean estimate, with error bars showing the interquartile range. Note that final size estimates are plotted on a log scale. The actual final outbreak size for each scenario is indicated by a horizontal dashed line.

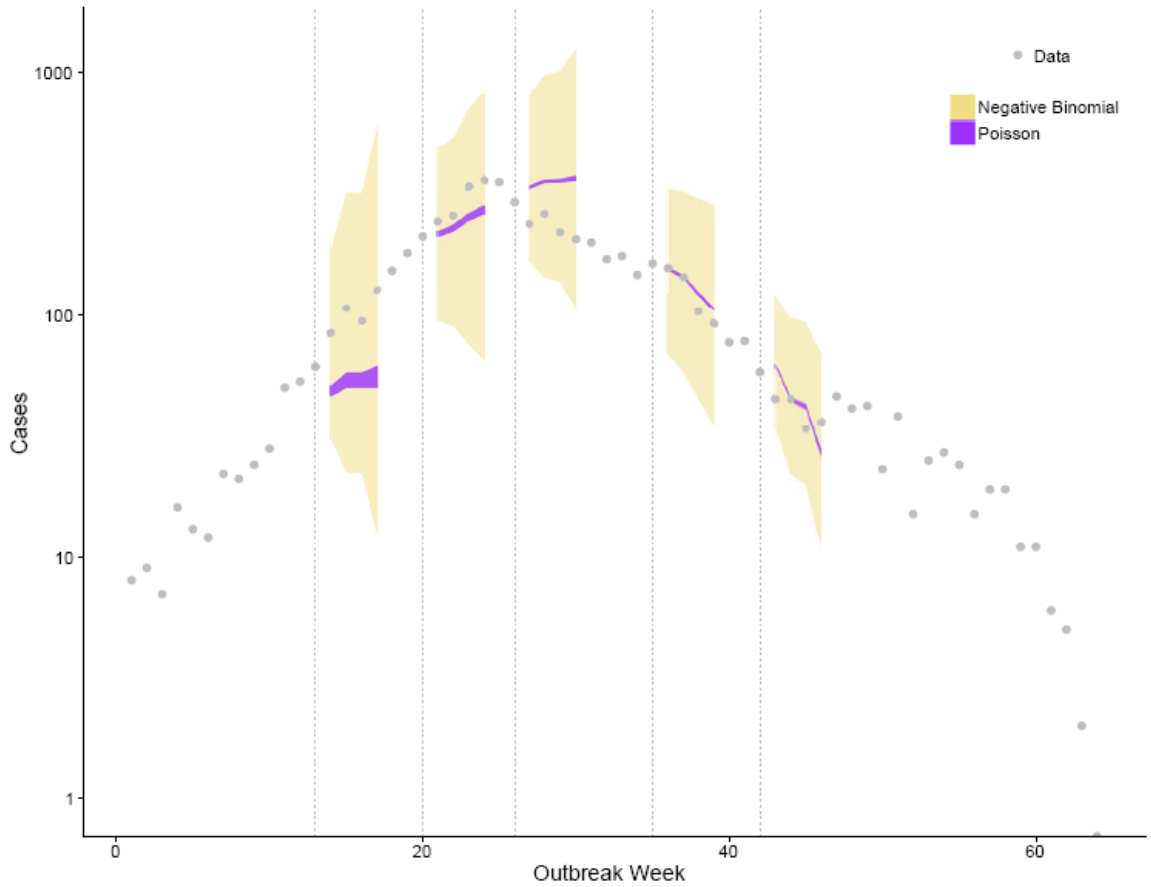

Figure 4. Comparison of negative binomial and Poisson models. We refit the model at each of the 5 time points for scenario 1, assuming a negative binomial distribution of cases, rather than the Poisson distribution used in the primary analysis. Circles represent provided data on weekly EVD cases. The interquartile range of projected weekly case counts is shown for the two models. Note that the cases are plotted on a log scale.

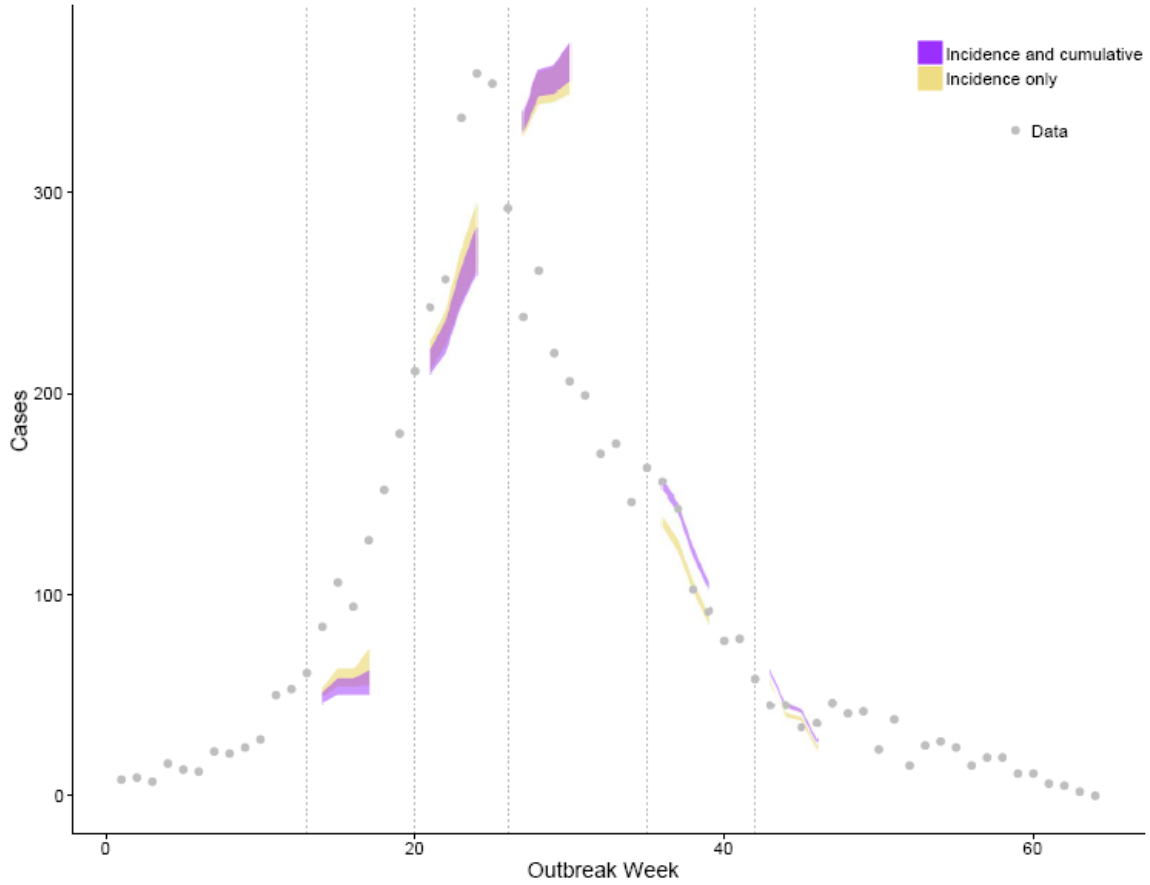

Figure 5. Difference in model projections when fitting to incident case data only. Short-term model projections are shown when the model was fit to incident and cumulative cases counts simultaneously (as in our main analysis) or to incident case counts only. Projections are shown for scenario 1, with evaluation time points indicated by dashed vertical lines. The interquartile range is shown for both approaches.