Epidemics 22 (2018) 50–55

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Forecasting Ebola with a regression transmission model

Jason Asher

Leidos Supporting the Department of Health and Human Services (HHS), Biomedical Advanced Research and Development Authority (BARDA), United States

ARTICLE INFO

Article history:
Received 1 July 2016
Received in revised form 4 January 2017
Accepted 17 February 2017
Available online 27 February 2017

Keywords:
Ebola
Forecasting
Mathematical modeling
Bayesian inference

ABSTRACT

We describe a relatively simple stochastic model of Ebola transmission that was used to produce forecasts with the lowest mean absolute error among Ebola Forecasting Challenge participants. The model enabled prediction of peak incidence, the timing of this peak, and final size of the outbreak. The underlying discrete-time compartmental model used a time-varying reproductive rate modeled as a multiplicative random walk driven by the number of infectious individuals. This structure generalizes traditional Susceptible-Infected-Recovered (SIR) disease modeling approaches and allows for the flexible consideration of outbreaks with complex trajectories of disease dynamics.

1. Introduction

The unparalleled scale of the 2014–2015 Ebola outbreak in West Africa and the lack of regulatorily-approved pharmaceuticals led to rapid acceleration in the development and clinical evaluation of therapeutics and vaccines for the prevention and treatment of Ebola Virus Disease. By the fall of 2014, two investigational vaccines, the GlaxoSmithKline (GSK) recombinant chimpanzee adenovirus 3 vector vaccine and the Newlink Genetics recombinant vesicular stomatitis virus vaccine, had already undergone Phase 1 clinical trials and initial results were promising (Rampling et al., 2016; Agnandji et al., 2016).

At that same time, Phase 2 and 3 clinical trials for vaccines were being concurrently planned in Liberia and Sierra Leone, and an additional study for therapeutics was being considered in Liberia and the United States (STRIVE, 2017; Henao-Restrepo et al., 2015). Determining the statistical power of these clinical trials – that is, their ability to detect whether the tested products were effective at preventing or treating infection – relied upon assumptions about the future incidence of Ebola in the geographic and demographic populations where the trials were to be conducted. As such, it became readily apparent that timely, regularly-updated, and accurate forecasts of the incidence of disease would be critical for ensuring that study plans were aligned with the realities of the rapidly-changing outbreak and for determining contingency trial planning scenarios.

While disease forecasting technologies were in demand for this and other practical applications, it appeared that future projections from traditional compartmental models systematically overestimated incidence and epidemic size (Bulter, 2014; Chretien et al., 2015). Furthermore, many published and publicized Ebola models either could not be easily adapted for iterative forecasting purposes, had too many parameters given the limited available data, or both. These models nonetheless provided important insights for controlling the outbreak, illustrating that if model assumptions were true, exponential growth in disease incidence could continue without a combination of significant external interventions, including increasing the number of Ebola Treatment Unit (ETU) beds available, effectively isolating patients earlier in their disease course, and changing burial practices (Gomes et al., 2014; Meltzer et al., 2014; Lewnard et al., 2014; Pandey et al., 2014).

In support of the U.S. Department of Health and Human Services Biomedical Advanced Research and Development Authority (BARDA), we developed an Ebola forecasting approach in order to assess the statistical power of the planned clinical trials (PHEMCE, 2017; HHRG, 2017). Our model assumed that multiple public health interventions would continue to be implemented and that our ability to delineate the relative contribution of these interventions to disease control would be limited. Rather than attempting to disentangle these mechanisms, our approach aimed to find a 'trend' in how these interventions and other general changes in public behavior might alter the overall rate of transmission. We then produced forecasts by projecting this trend into the future.

The initial forecasting methodology provided accurate probabilistic forecasts over a moderate term of approximately 2–3 months. In particular, this approach was able to accurately forecast future declines in incidence in Sierra Leone using data available in early November. These forecasts provided decision makers planning clinical trials with several months of advanced warning about

E-mail address: jason.m.asher@gmail.com

https://doi.org/10.1016/j.epidem.2017.02.009

the potential for these studies to be under-powered. Subsequently, the methodology was adapted to produce weekly iterative updates when new reported Ebola incidence data became available through WHO situation reports, even as these data were subject to regular modification and reclassification.

While successful in these contexts, the underlying model made strong assumptions – in particular, that the reproductive number decayed logistically over time – that were not appropriate for the epidemic trajectory in Guinea. These limitations left open questions about the applicability as a general Ebola forecasting approach in future outbreaks. As a submission to the NIH RAPIDD Ebola Forecasting Challenge, which took place September-December 2015, we adapted our initial forecasting methodology to build upon the lessons learned from our real-time operational Ebola forecasts. The Challenge brought together eight independent teams of modelers who strove to provide real-time forecasts of disease incidence across four synthetic Ebola outbreak scenarios in Liberia. The underlying data was generated by a spatially-explicit, stochastic, agent-based model of Ebola in Liberia that incorporated demographic effects and allowed for interventions such as changing burial practices and expanding Ebola Treatment Units. The scenarios varied in terms of data quality and quantity, with Scenario 1 providing the most high-quality data and providing access to synthetic outbreak trees and line lists. All scenarios incorporated noise and data loss to simulate real-world data availability to the extent possible.

Our model submission to the Challenge retained the reproductive number regression approach and parsimonious spirit of the original model, but made attempts to more accurately reflect the inherent uncertainty in future rates of transmission. In the text below we will detail the design and implementation of this model.

2. Methods

2.1. Model structure

The primary forecasting target for the Challenge was the prediction of the weekly incidence of Ebola across the four scenarios in Liberia. Secondary forecasting targets for prediction included the week and size of the peak (i.e. the greatest number of incident cases in a given week together with the week in which it occurs), and the final cumulative number of infections. In every scenario, and at every prediction point, the minimum guaranteed amount of information that would be available to produce the forecast would be the scenario-specific weekly incidence up to that point in time. The Ebola Forecasting Challenge description also indicated that additional situation report data describing the operation of ETUs and burial teams would be available (but subject to factual errors and omissions), and that at some points incomplete transmission trees would be distributed as well. Motivated by the belief that it would be challenging to directly quantitatively predict the effects of the situation report information on the simulated outbreaks, we opted to focus on using only the weekly incidence data as the input for our model.

We developed a discrete-time stochastic compartmental model with a time-varying reproductive number and a time step of 1 day (Fig. 1). The model had two discrete integer infection states – 'Exposed' and 'Infectious' with geometric transitions between compartments driven by a mean latent period and a mean infectious period that were assumed to be in the range of estimates from empirical studies (Table 1) (Van Kerkove et al., 2015).

The reproductive rate $r(t)$ was modeled as a multiplicative normal random walk with a log-linear drift driven by the current number of infectious individuals. Specifically,

$$\log(r(t)) = \log(r(t-1)) + \sigma Z(t) - \alpha I(t-1)$$

where $Z(t)$ is an independent and identically distributed sequence of standard normal random variables, σ is the dispersion of the random walk, α is the coefficient of drift, and $I(t)$ is the number of infectious individuals at time t. Given the model state at time t, new infections entering the 'Exposed' state were generated as a Poisson distribution with mean $r(t)I(t)$, and transitions between states were simulated in a first-order Euler fashion with independent geometric distributions at the corresponding rates. The use of a random walk for the reproductive rate was motivated by the successful application of similar approaches to modeling and forecasting Ebola (Camacho et al., 2015; Camacho et al., 2014; Shaman et al., 2014). The random walk also allows the model to follow complex incidence trajectories, including multiple 'waves' of infection seen in the reported data from Guinea during the 2014–2015 outbreak.

The particular log-linear drift used in this model generalizes traditional Susceptible-Infectious-Recovered (SIR) dynamics where the effective reproductive rate varies on a log scale proportionally to the number of currently-infected individuals (McKendrick, 1925). Indeed, in such a model, if one denotes by β the transmission rate, K the effective susceptible population, $S(t)$ the population susceptible at time t, $I(t)$ the population infectious at time t, and $r(t)$ the effective reproductive rate at time t, then one has

$$S'(t) = -\frac{\beta}{K} S(t) I(t)$$

and so

$$\log(r)'(t) = \frac{S'(t)}{S(t)} = -\frac{\beta}{K} I(t).$$

Models using SIR dynamics with smaller assumed effective susceptible populations have also been successfully used to reproduce the overall trajectories of the 2014–2015 Ebola outbreak in West Africa (Eisenberg et al., 2015). Additionally, the linear drift can also be considered as a term that implicitly incorporates behavioral change, modeled as a process that causes the effective reproductive rate to be decreased, declining proportionally to the current number of active infections.

2.2. Bayesian inference

The model was fit to the provided incidence data by applying Bayesian inference on the parameter and state space with assumed prior distributions in Table 1. In particular, we used ~30,000 iterations of a particle Markov-Chain Monte Carlo (MCMC) method that combines Metropolis-Hastings updating in parameter space with Sequential Monte Carlo state-space sampling (Andrieu et al., 2010). Parameters were perturbed by independent Gaussian distributions whose variances were tuned by inspection to ensure that individual parameter samples were sufficiently dispersed and that the overall acceptance rate was approximately 30 percent, consistent with literature on the optimal tuning of such algorithms (Gelman et al., 1997). The resulting samples from the posterior distribution were in turn used to produce credible intervals for the historical and predicted future states, including the reproductive rate and incidence. All of this was implemented in R and C++ using the Rcpp package for computational efficiency, and charts were similarly produced in R using the ggplot2 package (R Core Team, 2014; Eddelbuettel and Francois, 2011; Wickham, 2009). The entire workload to produce forecasts for all four scenarios at a given prediction time point required less than 1 h using a single core from a modern workstation desktop computer, including computation, inspection of results, and post-processing.

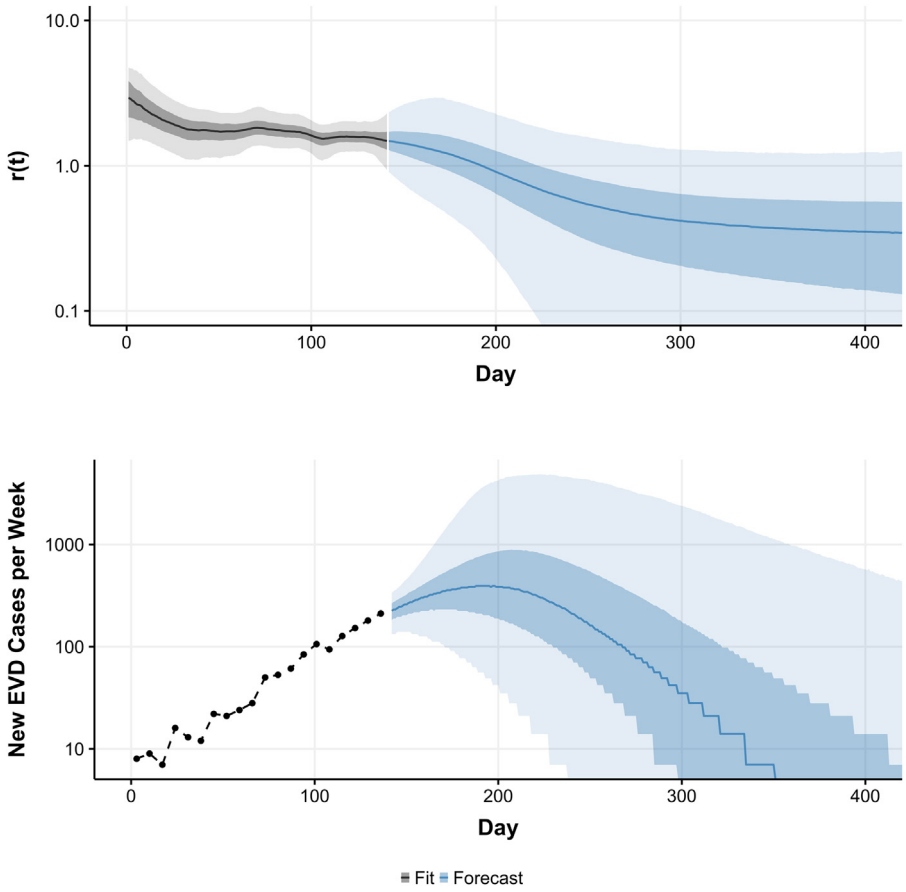

Fig. 1. Illustration of fit and forecast for Scenario 1, Prediction 2. Input incidence data is denoted by black dots and dashed lines in the lower left. The resulting estimated reproductive number during the same period is denoted with black lines and shaded regions in the upper left. This reproductive number, together with the resulting incident infections, is in turn forecasted by the model as illustrated with blue lines and shaded regions. In all shaded regions, the central line indicates the median, the darker shaded region indicates the interquartile range, and the lighter shaded region indicates the 5–95th percentile range.

Table 1
Parameters and prior distributions used for Bayesian inference.

Parameter	Prior distribution	Notes and source
$r(0)$	Uniform on [1, 5]	Assumption based upon prior range of Ebola estimates (Van Kerkove et al., 2015)
$r(t)$ random walk dispersion – σ	Log-uniform on (0,1]	Vague prior – upper bound chosen arbitrarily and large enough so that higher values would be highly improbable.
$r(t)$ random walk drift coefficient – α	Uniform on [0,K] $K = \frac{1000}{4.4e6*8}$	Vague prior – upper bound chosen to allow up to ~1000× faster drift than in SIR dynamics, which is equivalent to assuming an effective population fraction of 0.001 as in (Eisenberg et al., 2015). Indeed, assuming $r(0) \sim 1$–2, an infectious period of ~8–16 days, and a population of ~4.4e6, SIR drift would be $\sim \frac{1}{4.4e6*8}$.
Latent period	Uniform on [8, 12] days	Assumption based upon prior range of Ebola estimates (Van Kerkove et al., 2015)
Infectious period	Uniform on [8, 16] days	Assumption based upon prior range of Ebola estimates (Van Kerkove et al., 2015)
Initial infections	Uniform on [1, 10]	Assumption based upon initial inspection of data – upper bound chosen arbitrarily for properness of prior.

2.3. Model diagnostics

Model performance was assessed as part of the Challenge using Pearson's r and the coefficient of determination to analyze correlation between median forecasted and actual incidence, as well as using mean absolute error, mean squared error, and mean absolute percentage error. Furthermore, to assess the estimation of uncertainty in forecast and predicted metric distributions, for each scenario, prediction we calculated the percentage of predictions that contained the true value in the inter-quartile range and in the 5–95th percentile ranges. If the interquartile and 5–95th percentile ranges contained approximately 50% and 90% of the true values respectively, we interpreted that the ranges of the forecast distributions were appropriately calibrated.

3. Results and discussion

We developed a state-space inferential framework with a semi-mechanistic discrete-time stochastic compartmental disease model that generally produced accurate, moderate-term (2–3

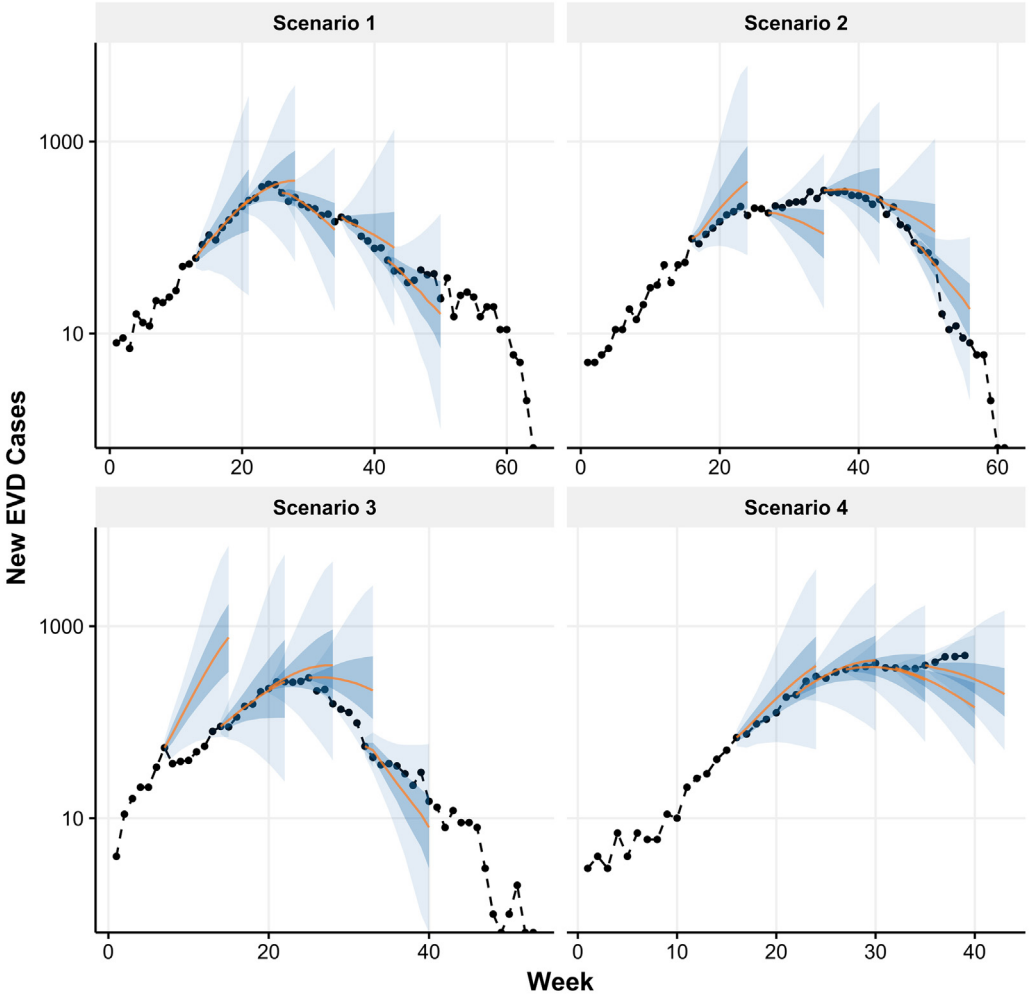

Fig. 2. Ebola Virus Disease (EVD) incidence forecasts for each prediction time point. Scenario incidence data is in black. Central orange lines indicate median incidence forecast. Darker shaded region indicates forecast interquartile range, and lighter shaded region indicates 5–95th percentile range. All displayed forecast durations are 8 weeks from point of prediction.

months) incidence forecasts across scenarios, as illustrated in Fig. 2. By design in the Challenge, Scenario 1 had the highest-quality and most accurate data, and the corresponding forecasts were highly accurate with median predictions very close to observed values and with the interquartile ranges of the forecasts produced at four of five prediction points capturing all of the scenario incidence data.

In Scenario 2, at prediction point 2, the model suggested that incidence may have already peaked, but the predictions were highly uncertain and the upper end of the range provided for the possibility of further growth (as actually occurred). For the remaining prediction points, the forecasted values agreed largely with the future scenario data.

In Scenario 3, the initial time point prediction suggested more rapid growth in incidence than what was later observed, but that prediction time point was likely selected specifically for this effect–the data available up to that indeed seemed to suggest growth as the model predicted. Fits obtained in the subsequent weeks would be more in line with the actual observed future number of cases. Additionally, the remaining forecasts were more well-aligned with the future outbreak trajectory.

Finally, in Scenario 4, the model continued to suggest that incidence may have already peaked by the last prediction point, but the uncertainty range still indicated the possibility of future growth in incidence. This was aligned with reality where, in truth, there was no peak in sight at the end of the simulation period for Scenario 4. Furthermore, the early incidence forecasts were highly accurate, with all of the true values lying inside the predicted interquartile range and close to the median predictions.

The incidence forecasts indicated a large degree of uncertainty, with interquartile ranges spanning approximately a 3- to 4-fold range after 8 weeks and 5–95th percentiles spanning approximately a 30- to 50-fold range after 8 weeks. But, much of this uncertainty was warranted as the comparison of model probabilistic calibration indicated. Indeed, out of the 160 forecasts made across scenarios and 1–8 weeks after each prediction time points, 64% of incidence predictions contained the true value in the interquartile range, and 96% of incidence predictions contained the true value in the 5–95th percentile ranges, suggesting only moderate over-estimation of the actual uncertainty in the predictions. Additionally, further tuning of the prior distributions, in particular the dispersion in the reproductive rate random walk, should be able to control this over-estimation.

As Fig. 3 depicts, the model was also able to provide one month advance warning of the timing of peaks in incidence as well as the sizes of those peak at prediction points. Across the three targets – peak week, peak size, and final epidemic size – the predicted values converged to the actual values as more data was available (i.e. the distributions of the ratio of predicted estimate to actual data converged to 1). Furthermore, the model produced reasonably accurate final outbreak size estimates 30–40 weeks in advance

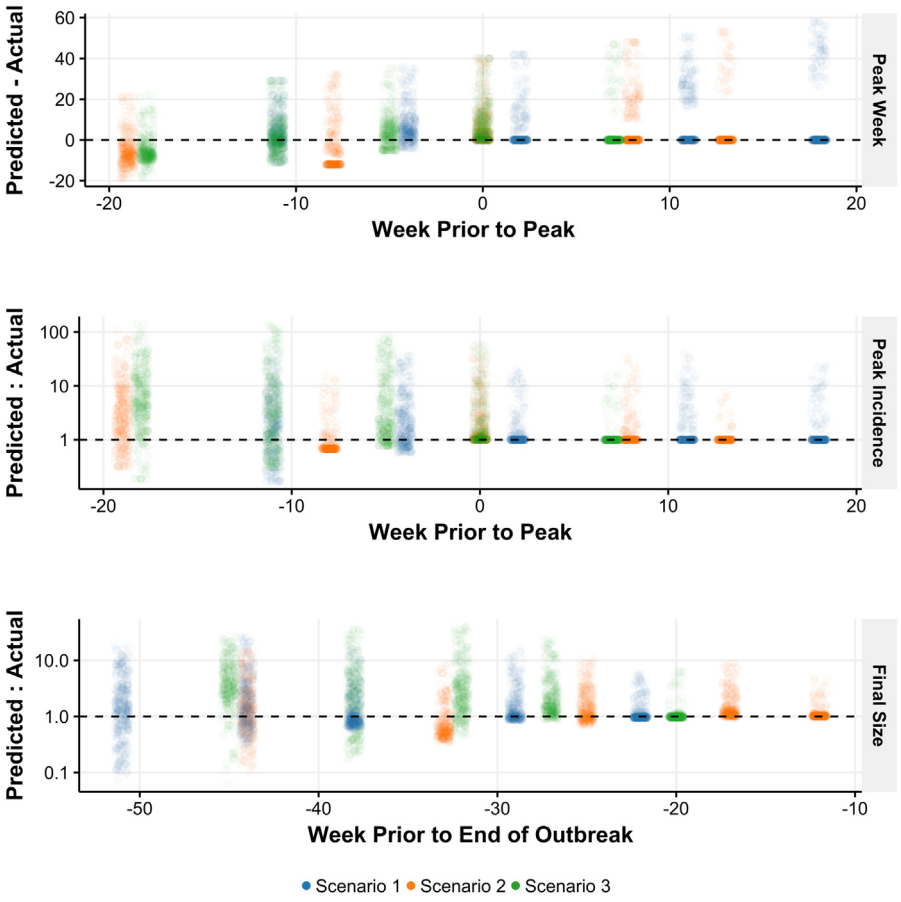

Fig. 3. Difference of predicted and actual peak week (top), ratio of predicted to actual peak incidence (middle), and ratio of predicted to actual final outbreak size (bottom) across scenarios and prediction time points. Plotted points represent ∼300 samples from the posterior distribution for the relevant metric with a random jitter applied to the time coordinate to show detail and improve visibility. For metrics associated with a peak in incidence, time is given in weeks prior to this event. For the final size, time is given in weeks prior to end of the outbreak. For both the final size and peak metrics, ratios are presented on a log scale. Some relative prediction times overlap across scenarios. Scenario 4 is not included because the comparative events did not occur during the simulation period of the Challenge. The model structure allowed for the possibility of later, larger peaks (and larger overall outbreaks) even after the scenario peak had been largely predicted to have already occurred–this explains the presence of some samples higher than the true values in later predictions.

– in some cases before the peak – with increasingly accurate estimates as time passed. As with the incidence estimates, the range of uncertainty was very large with interquartile ranges for the size of a peak a month in advance being roughly 6- to 8-fold. Peak estimates were reasonably well-calibrated with 57% of actual peak and peak week data lying within the interquartile range of predictions made prior to the peak. Final size predictions were slightly worse in terms of calibration with 71% of such inter-quartile range predictions containing the true values. Additionally, when the ranges did not contain the true value, the cause was invariably over-estimation.

Despite this generally good performance, it is important to note that the underlying model still only represents an initial prototype technology and needs a good deal of refinement and further experimentation. Using a random walk for the effective reproductive number has the effect of allowing underlying model misspecification to be compensated for by large random walk dispersion. Thus, we had to constrain this parameter to a certain extent, and the overall relationship between the choice of prior distributions and resulting model performance still needs to be explored. Furthermore, parameter inference on the coefficient of drift in the random walk was not powerful, and there seemed to be an inferential tension between power of inference on this parameter, and the overall random walk drift. It remains to be seen if a better choice of prior distribution will improve inferential, and in turn forecast, performance.

The strength of this forecasting model was that it was relatively simple to implement and efficient to utilize, with the entire computational and labor time required to produce forecasts being less than an hour for all four scenarios. The inference methodology was computationally and numerically robust and required little effort to inspect and tune. The overall model incidence forecast performance was strong with the lowest overall mean absolute error and mean squared error and highest Pearson's correlation among models in the Challenge. Additionally, it had the second-lowest mean absolute percentage error and second-highest overall coefficient of determination. Finally, although it was only used to produce national-level forecasts, this methodology can be readily extended to consider district-level data and include the resulting heterogeneities in rates of transmission.

The underlying model design shared a great deal with the above-mentioned approach taken to produce incidence forecasts for BARDA during the Ebola outbreak in West Africa. It similarly strove to minimize the number of internal degrees of freedom so as to maximize the inferential value of limited data. It also similarly fell into a class of 'regression transmission models' that attempted to consider outbreak forecasting as a problem of fitting an appropriate statistical model to the time-series of the reproductive number as a function of time. This class of models can include versions with more explicit representations of mitigation methods, for example by including a time-series of ETU bed availability as a covariate, or

adding the time since relevant events like major changes in policy or case count threshold. Thus, it seems like a natural family of relatively simple candidate forecasting models that are worth of future study.

4. Conclusion

Our work has demonstrated that a relatively simple semi-mechanistic mathematical model of Ebola transmission can be used to produce accurate forecasts of the incidence of Ebola at the national level that anticipate future peaks using only weekly incidence data. During the 2014–2015 Ebola outbreak in West Africa, WHO did not begin publishing weekly incidence until mid-November 2014, and the data that was available was subject to significant delays in reporting and revisions. This was understandable, certainly, given the extraordinary circumstances surrounding the outbreak and the need to prioritize resources toward treatment and containment, but this study illustrates the potential decision-making value that could be obtained if timely and accurate incidence data can be made available publicly as early as possible during an outbreak.

The underlying agent-based model that was used to generate the forecast challenge data included explicit changes in behavior over time. For diseases like Ebola, i.e. those with high consequences of infection and highly visible symptoms, behavioral change will likely be a major factor in the progression of an outbreak. Modeling these behavioral changes is a difficult endeavor, but may be a critical component of accurate disease forecasting in future work (Funk et al., 2015). Large-scale simulation models are one possible approach, but these can struggle with the lack of sufficiently detailed input data and with the difficulty of matching the specific trajectory of a given outbreak given the huge number of internal degrees of freedom. Our approach favored simplicity and was successful, but it still leaves much to be improved upon. Nonetheless, decision makers have an appetite and genuine need for these forecasts and the modeling community should continue the work started during the outbreak, and continued with this Challenge, to further develop and refine the requisite technology.

References

Agnandji, S., Huttner, A., Zinser, M., et al., 2016. Phase 1 trials of rVSV Ebola vaccine in Africa and Europe. N. Engl. J. Med., http://dx.doi.org/10.1056/NEJMoa1502924 (preliminary report published online April 1, 2015).

Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle Markov chain Monte Carlo methods. J. R. Stat. Soc. B 72, 269–342.

Camacho, A., Kucharski, A., Aki-Sawyerr, Y., 2015. Temporal changes in ebola transmission in sierra leone and implications for control requirements: a real-time modelling study. PLoS Curr. Outbreaks, http://dx.doi.org/10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2.

Bulter, D., 2014. Models overestimate Ebola cases. Nature 515, 7525.

Camacho, A., Kucharski, A.J., Funk, S., Breman, J., Piot, P., Edmunds, W.J., 2014. Potential for large outbreaks of Ebola virus disease. Epidemics, http://dx.doi.org/10.1016/j.epidem.2014.09.003.

Chretien, J.P., Riley, S., George, D.B., 2015. Mathematical modeling of the West Africa Ebola epidemic. eLife, http://dx.doi.org/10.7554/eLife.09186.

Eddelbuettel, D., Francois, R., 2011. Rcpp: Seamless R and C++ Integration. J. Stat. Softw. 40 (8), 1–18, http://www.jstatsoft.org/v40/i08/.

Eisenberg, M., Eisenberg, J., D'Silva, J.P., et al., 2015. Forecasting and Uncertainty in Modeling the 2014–2015 Ebola Epidemic in West Africa (Preprint) arXiv: 1501.05555.

Funk, S., Bansal, S., Bauch, C.T., et al., 2015. Nine challenges in incorporating the dynamics of behaviour in infectious diseases models. Epidemics, http://dx.doi.org/10.1016/j.epidem.2014.09.005.

Gelman, A., Gilks, W.R., Roberts, G.O., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. 7 (1), 110–120.

Gomes, M.F.C., Pastore y Piontti, A., Rossi, L., Chao, D., Longini, I., Halloran, M.E., Vespignani, A., 2014. Past Projections, October 6, http://www.mobs-lab.org/ebola.html.

Henao-Restrepo, A.M.H., Longini, I.M., Egger, M., 2015. Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial. Lancet, http://dx.doi.org/10.1016/S0140-6736(15)61117-5.

HHRG, Examining the U.S. Public Health Response to the Ebola Outbreak, 2017. http://docs.house.gov/meetings/IF/IF02/20141016/102718/HHRG-113-IF02-Wstate-RobinsonR-20141016.pdf.

Lewnard, J.A., Ndeffo Mbah, M.L., Alfaro-Murillo, J.A., Altice, F.L., Bawo, L., Nyenswah, T.G., Galvani, A.P., 2014. Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis. Lancet Infect. Dis. 14, 1189–1195.

McKendrick, A.G., 1925. Applications of mathematics to medical problems. Proc. Edinburgh Math. Soc. 44, 98–130, http://dx.doi.org/10.1017/S0013091500034428.

Meltzer, I., Atkins, C.Y., Santibanez, S., et al., 2014. Estimating the future number of cases in the Ebola epidemic – Liberia and Sierra Leone, 2014–2015. MMWR 63 (Suppl. 3), 1–14.

Pandey, A., Atkins, K.E., Medlock, J., et al., 2014. Strategies for containing Ebola in West Africa. Science 346, 991–995.

Public Health Emergency Medical Countermeasures Enterprise (PHEMCE), 2017. Strategy and Implementation Plan. http://www.phe.gov/Preparedness/mcm/phemce/Documents/2015-PHEMCE-SIP.pdf.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rampling, T., Ewer, K., Bowyer, G., et al., 2016. A monovalent chimpanzee adenovirus Ebola vaccine boosted with MVA. N. Engl. J. Med., http://dx.doi.org/10.1056/NEJMoa1411627 (preliminary report published online January 28, 2015).

Shaman, J., Yang, W., Kandula, S., 2014. Inference and forecast of the current west African Ebola outbreak in Guinea, Sierra Leone and Liberia. PLoS Curr. Outbreaks, http://dx.doi.org/10.1371/currents.outbreaks.3408774290b1a0f2dd7cae877c8b8ff6.

Sierra Leone Trial to Introduce a Vaccine against Ebola (STRIVE), 2017. http://www.cdc.gov/vhf/ebola/strive/index.html.

Van Kerkove, M.D., Bento, A.I., Ferguson, N.M., Donnelly, C.A., 2015. A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. Sci. Data, http://dx.doi.org/10.1038/sdata.2015.19.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.