An IDEA for Short Term Ebola Outbreak Projection

ELEYINE ZAROUR

ABSTRACT

Fisman et al have described a simple two-parameter mathematical model for short term outbreak projection that provided an accurate assessment of pandemic influenza A (H1N1) outbreak test data [1]. In this paper, we will apply this model to the most recent Ebola epidemic growth patterns in West Africa.

1 INTRODUCTION

In the face of novel pathogens such as the SARS coronavirus or genetically diverse familiar pathogens such as the Ebola virus, public health authorities are confronted with the practical task of outbreak management. To inform public health policies, they often resort to mathematical models to analyze the spread and control of infectious diseases. Typical models – most notably the Susceptible-Infected-Removed (SIR) model and its derivatives – represent epidemics as processes resulting from the transition of individuals between health states.

However, in the early context of an outbreak where only limited data is available, the correct parametrization of such "explicit" models is difficult as it requires detailed information on incidence, immune status in the population and contact patterns. The latter is particularly difficult to estimate given the transmission-reducing behavioral change of the population following intervention measures and media coverage [3] and the effect of imported cases in an increasingly globalized world [6, 8]. Moreover, such models fail to account for the heterogeneity of the affected population which consists of subpopulations with different transmission rates (e.g. influenza initated among children) [4]. In addition, the genetic diversity of the pathogen itself may result in clade-specific growth rates and reproduction numbers. For example, recent studies have identified new mutations in the Sierra Leone Ebola outbreak that, should they remain stable, will generate major shifts in clade frequencies and influence the overall epidemic dynamics on time scales within the current outbreak [5, 7]. Finally, early outbreak incidences are often unreported or misdiagnosed due to the unavailability of microbiological or serological diagnostic tools.

The time-dependent effects of these various factors on the effective reproduction number R are difficult to quantify. The Incidence Decay and Exponential Adjustment (IDEA)

[1, 2] model proposed by Fisman et al. is suitable as a descriptive and prognostic tool for early epidemic outbreaks because it requires limited data by design and empirically takes into account the dampening effect due to control intervention. In addition to assessing the degree of current control efforts, the IDEA model is robust to noise and provides credible and easily interpreted projections on epidemic size and duration.

2 MODEL

2.1 Model Description

The IDEA model requires two input parameters: the daily incidence count (I) and the average serial interval (t) which is the time between symptoms developing in an index case and symptoms developing in a secondary case. The serial interval is symptom-based so as to be applicable in early outbreak stages when microbiological or serological diagnosis is not yet available.

The data is then used to fit two parameters characteristic of the outbreak's epidemic growth: the basic reproduction number R_0 and the dampening factor d. R_0 is defined by Vynnycky and White as "the (average) number of successful transmissions per infected person" when an infected person first enters a completely susceptible population. As an outbreak progresses, the effective reproduction number is reduced due to the depletion of susceptible individuals but also due to several other factors outlined in Section 1, most notably public health intervention and transmission-reducing behavioral changes. The impact of this time-dependent dampening factor can be expressed empirically as:

$$I(t) = \left[\frac{R_0}{(1+d)^t}\right]^t \tag{1}$$

Thus, when there is no control and $d = 0$, the equation reverts to early-stage characteristic exponential growth $I(t) = R_0^t$. When $d > 0$, this exponential growth is reduced by a factor of t^2, causing transmission to slow and stop even when d is small.[1]

2.2 Model Application

Best-fit parameters R_0 and d are estimated by minimizing the root mean-squared distance (RMSD) between model estimates and empirical data.

In addition to generating estimates of R_0 and d, the outbreak duration t_max can be estimated by solving $I(t) < 1$ in (1) such that:

$$t_{max} \geq \frac{\ln R_0}{\ln(1+d)} \tag{2}$$

1 The authors make a mistake in the manuscript by stating that transmission slows and stops "even when the absolute value of d is small". On the contrary, d < 0 would result in even more pronounced exponential growth.

Moreover, integrating (1) over t provides an expression to estimate the total outbreak size I_{total} (Equation (5)).

Finally, though the structure of the IDEA model makes it impossible to fit multi-wave epidemics, an increasing Δd is an important indicator of the emergence of a new wave of infection as demonstrated in Appendix A.

$$\Delta d = d_i - d_{i-1} \tag{3}$$

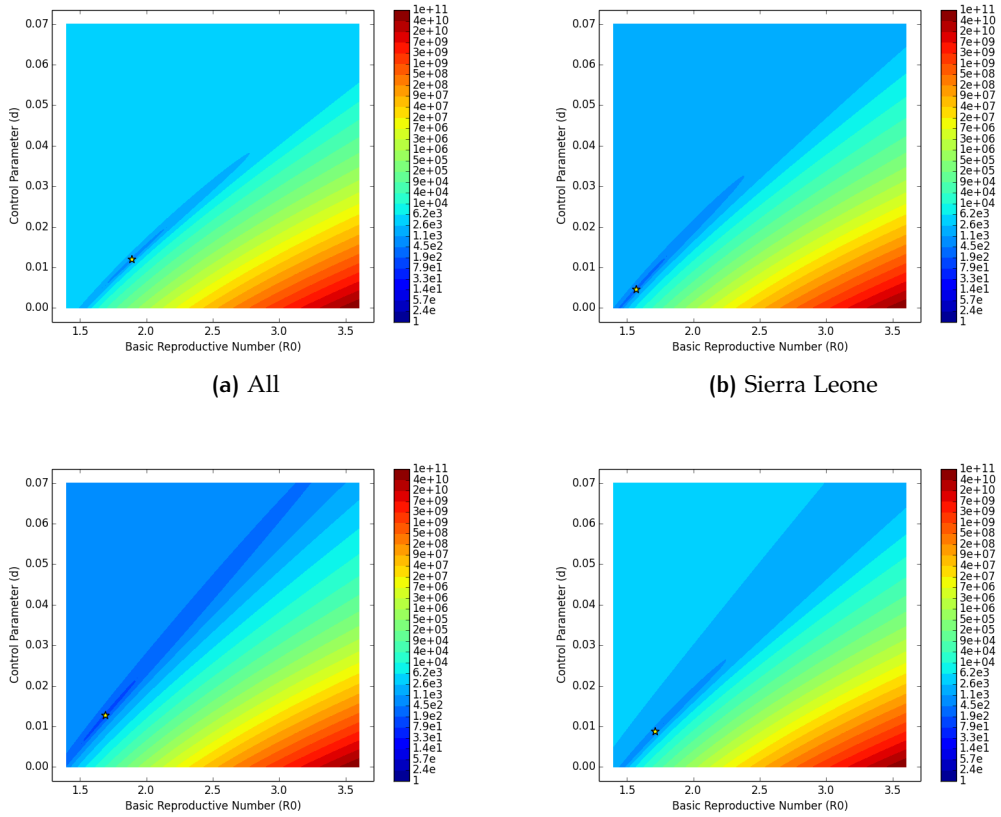
2.3 Model Assessment

To assess the model's performance, we simulate epidemics using a discrete SIR model under different assumptions about infectiousness and varying orders of control. We then fit the IDEA model to the simulated data by minimizing the RMSD between generation-specific case counts and compare the fitted model parameters $R0$ and d to the original simulation parameters. The IDEA model was found to work best for first order control diseases with low R_0. More details on this method can be found in Appendix A .

3 RESULTS

3.1 Model Fitting

IDEA Model fits were performed on the most recent data set by minimising RMSD (Figure 1, Table 1). RMSD of the overall cumulative incidence data were lowest by an order of mangitude for R_0 values close to 1.89 and 0.01. These values are consistent with the values obtained by Fisman et al using earlier epidemic data.



(a) All **(b)** Sierra Leone

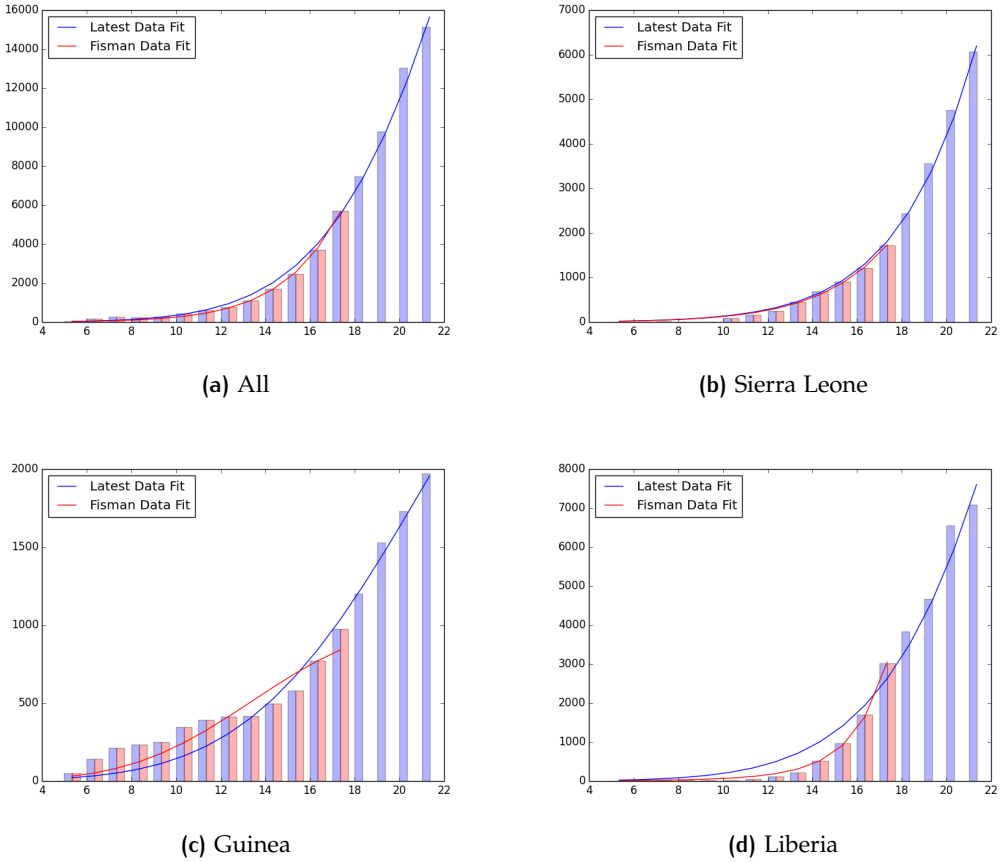
(c) Guinea **(d)** Liberia

Figure 1: Contour plot of RMSD between Observed Data and IDEA Model for given R_0 and d. Stars indicate the best-fit R_0 and d. Asssumes 15 day serial interval and first cases have been reported in generation 5.

Table 1: Best-fit Ro and d by Country

	All Countries	Guinea	Liberia	SierraLeone
Ro	1.892	1.693	1.714	1.571
d	0.012	0.013	0.009	0.005

The best fit models for all countries are plotted in Figure 2 and summarized in Table 2. There is a pronounced difference in model fit parameters between the Fisman Dataset (up till August 2014) and our updated dataset (up till November 2014) for Guinea and Liberia. These differences are likely due to higher sensitivity arising from these countries' relatively low incidence count. Thus, Fisman et al's model fits have overestimated and underestimated outbreak growth for Liberia and Guinea respectively.



(a) All

(b) Sierra Leone

(c) Guinea

(d) Liberia

Figure 2: Observed vs Fitted Cumulative Incidence Count by Country. Asssumes 15 day serial interval and first reported cases have been reported in generation 5.

Table 2: Caption

	All Countries	Guinea	Liberia	SierraLeone
Ro	1.892	1.693	1.714	1.571
Fisman Ro	1.721	2.007	1.232	1.545
d	0.012	0.013	0.009	0.005
Fisman d	0.005	0.027	-0.014	0.004
RMSD	280.987	103.328	337.097	80.471
Fisman RMSD	107.435	91.122	48.350	48.218

3.2 Projections

We computed the estimated outbreak size and duration based on progressively increasing numbers of epidemic generations (Figure 3). As of now, the outbreak is predicted to last 61 generations (of 15 days each) and affect 198,718 individuals in total. However, fitting

the data on Sierra Leone alone gives much greater (unrealistic) estimates ($I_{total} = 127\ 899\ 426$, $t_{max} = 142$). This is due to the model's high sensitivity as discussed in Section 3.5.

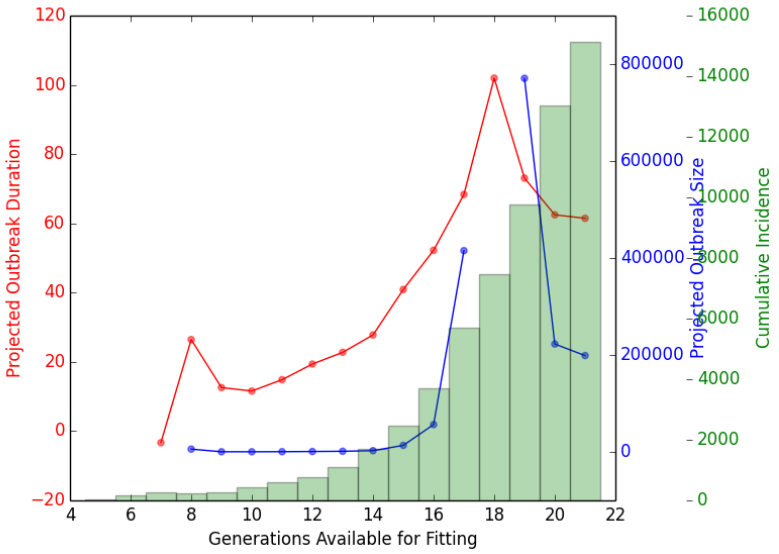
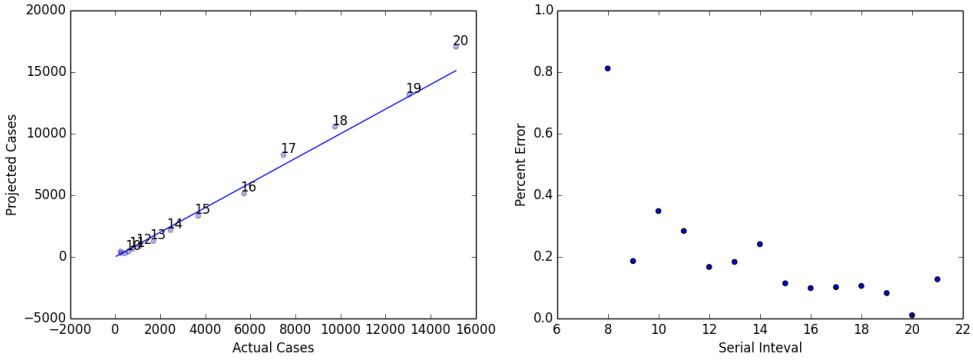


Figure 3: Projected Outbreak Size and Duration based on Guinea, Liberia and Sierra Leone cumulative incidence data.

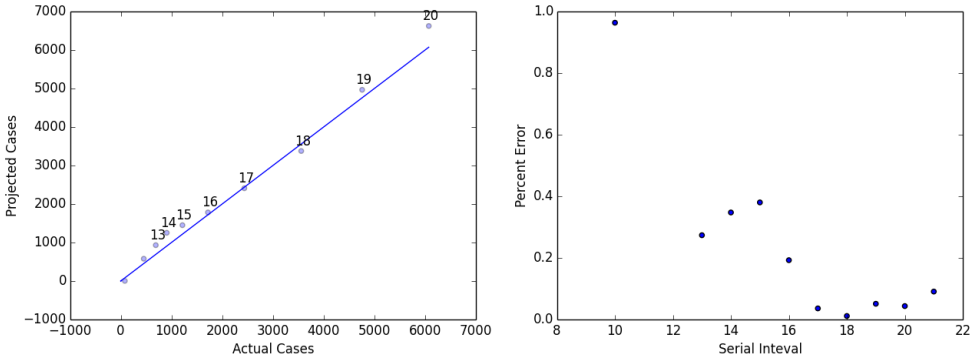
3.3 Intervention Control Assessment

The IDEA model is particularly useful in evaluating the impact of public health and social or environment factors on outbreak behavior. Thus, by projecting incidence count to the $(i+1)^{th}$ interval based on the oubreak up to i intervals and comparing the projection to the actual cases, it is possible to assess whether the outbreak is under control. For y-values superior to $y = x$, the projection overestimated the severity of the outbreak and therefore control measures performed better than expected.

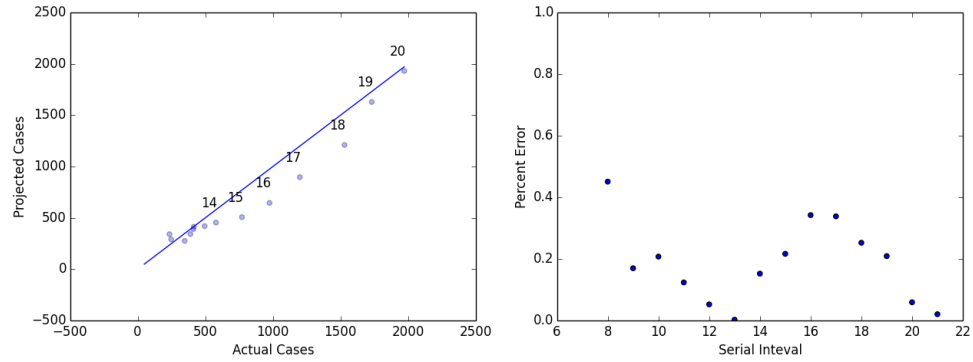
According to the projections outlined in Figure 4, Liberia cases are consistently overestimated which is indicative of a good level of control whereas Guinea projections follow the opposite trend. This contasts with Fisman et al's claim that "the threat [of Ebola] is currently centerd on the Liberian component of the epidemic which can be characterized as a simple exponential growth process, with little evidence for slowing of transmission". Though the Liberian outbreak does show exponential growth, it remains less pronounced than the Sierra Leone outbreak which demonstrated a misleading short-lived slowing of transmission around the publication of Fisman et al's paper.



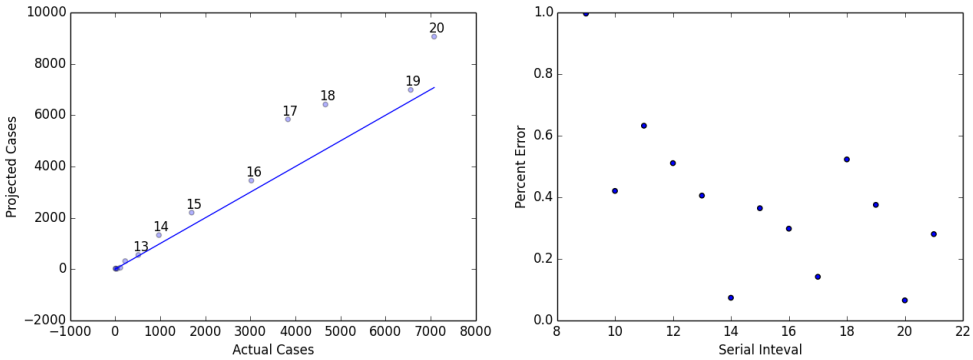
(a) All



(b) Sierra Leone



(c) Guinea



(d) Liberia

Figure 4: Projected vs Actual Cases Incidence counts to the $(i+1)^th$ interval are projected based on the oubreak up to i intervals. Percent error between projected and actual values are also plotted.

3.4 Multi-wave Epidemics

Post-hoc, it is trivial to recognize that Sierra Leone's outbreak has undergone sequential "waves" either due to the impact of seasonal or behavioural influences on disease tramsission, the movement of epidemics into previously unaffected sub-populations or a failure of control measures.

As mentioned in Section 2.2, the authors argue that Δd is a useful metric to monitor multi-wave epidemics whereby large positive Δd might predict the onset of a new "wave". However, there does not seem to be a noticeable increase in Δd at generations 16-17 – the last generations analyzed by Fisman et al – which weakens their claim (Figure 5) .

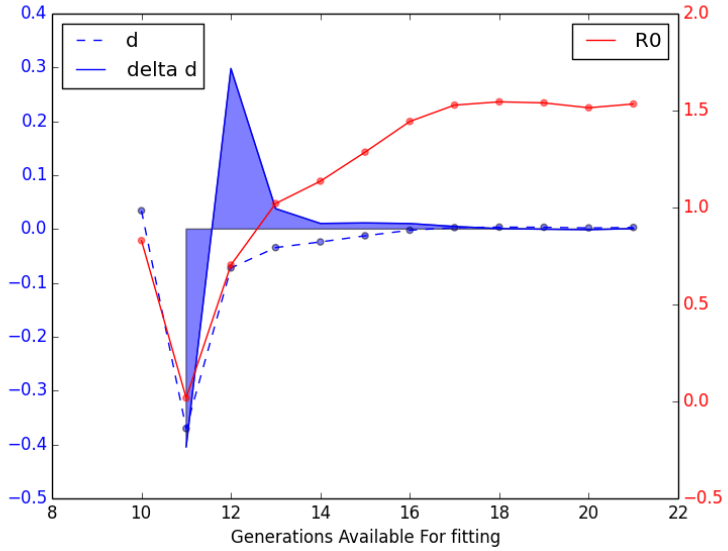


Figure 5: Sierra Leone Δd and best-fit IDEA parameters over time

3.5 Sensitivity Analysis

To evaluate the sensitivity of model estimates, we fit IDEA model parameters under alternate approaches and assumptions (Table 3). We compute these parameters for both our dataset and the earlier epidemic dataset used by Fisman. Though we were unable to precisely reproduce the parameters obtained by Fisman et al, the values remain close.[2]

According to the authors, the model fit is robust to case under-reporting. However, we observed a significant change in R_0 and d due to under-reporting. To simulat under-reporting at a ratio p, we simply scaled down the cumulative incidence data with a factor of p. The discrepancy between our results could be explained by the use of another method by the authors to downsample case reports.

Sensitivity data for individual countries are also available in the Appendix.

2 Fisman et al examine the peculiar assumption where "100% of the outbreak is under-reported", I was not sure how to compute that since the cumulative count would be 0 at all times... Not sure what's going on there. I tried 99% instead.

Table 3: Sensitivity Analysis on All Countries

	Ro	Ro (Fisman)	d	d (Fisman)
Base Case	1.892	1.721	0.012	0.005
12 day generation time	1.752	1.583	0.010	0.004
18 day generation time	2.181	1.863	0.018	0.007
Outbreak recognized generation 3	2.171	2.049	0.019	0.014
Outbreak recognized generation 7	1.697	1.510	0.007	0.000
Outbreak 50% underreported	4.293	3.682	0.061	0.039
Outbreak 99% underreported	4.010	3.591	-1.000	-1.000
Deaths only	1.852	1.646	0.014	0.006

4 DISCUSSION

Overall, model fits to most recent cumulative incidence data (up to November 2014) corroborate the authors' findings based on older data (up to August 2014). Like them, we found that the outbreak in West Africa (in Sierra Leone, Liberia and Guinea) was characterized by an $R_0 \approx 1.8$ and a a low dampening factor $d \approx 0.01$.

When inspecting the results specific to each country however, we note significant discrepancies between our findings. While the authors claim that the most worrisome component of the epidemic is Liberian, the findings detailed in this paper suggest that Sierra Leone's lack of intervention control is more threatening.

Nonetheless, both of our results point to exponential-like epidemic growth patterns in West Africa with little evidence of slowing of transmission (small d). It is not possible to attribute explicit mechanisms to the dampening factor d but the fact that Guinea's incidence data is consistently characterized by a relatively superior d seems to suggest that an increase in control is attainable for Liberia and Sierra Leone.

5 MATERIALS AND METHODS

5.1 Code

All computations and figures used in this paper were performed using the Python libraries `Numpy` (Numerical Python), `Scipy` (Scientific Python) and `Matplotlib` (Python Mathematical Plotting Library). The source code is available on Github at `https://github.com/eleyine/EbolaIDEAModel`. In particular, the file `paper.py` is structured in the same order as this paper and contains separate routines for each plot.

5.2 Sample Data

The analysis has been performed using WHO data as of November 30th, graciously compiled and published on Github by Ms C. Rivers. The data can be found at `https://github.com/cmrivers/ebola/`.

5.3 Serial Interval

We use the following serial interval heuristic and assume that incubation is equivalent to latency for the Ebola virus.

$$t = \text{incubation} + \frac{1}{2}\text{infective period}$$

$$\begin{cases} \text{incubation period} \approx 13\text{days} \\ \text{infectivity} = [3,5]\text{days} \end{cases} \rightarrow t \in [12,18]\text{days}$$

We use $SI = 15$ days throughout the analysis but vary the serial interval to $SI = 12$ and $SI = 18$ in the sensitivity analysis (Section 3.5).

5.4 SIR Model Comparative Assessment

To evaluate the IDEA model performance, Fisman et al simulate epidemics using a SIR discrete model where:

$$\begin{aligned} S_{t+1} &= S_t - \mathbf{R}e_t I_t \\ I_{t+1} &= \mathbf{R}e_t I_t \\ R_{t+1} &= R_t + I_t \\ N &= S + I + R \end{aligned} \tag{4}$$

Note: **R** is the "Reproductive" number (the average number of successful transmissions per infected person) and R is the number "Removed" individuals (I didn't come up with the notation.)

In the above SIR model, $\mathbf{R}e_t$ accounts for control activities and dynamic changes in population behavior that may reduce transmissibility of infection where:

$$e_t = RR^{t^n}$$

such that RR = relative risk of transmission and n = "order" of control
Thus,

For $n = 0$, $e_t = RR$ and Re is simply reduced by a constant fraction throughout the epidemic.

For $n = 1$, $e_t = RR^t$ and Re is reduced in a manner that accelerates with time.

For $n = 2$, $e_t = RR^{t^2}$ represents accelerated acceleration of control etc.

Though I successfully implemented the discrete SIR model (see `get_SIR` in `functions.py`), I did not manage to find the values the authors used for RR and N to generate testing data similar to theirs and thus reproduce their results.

5.5 Data Extrapolation

Local linear extrapolation was performed on missing daily incidence count data points (Figure 6). This extrapolation is important to obtain the cumulative incidence count for all countries studied so that data points from each country can be summed at any given time.

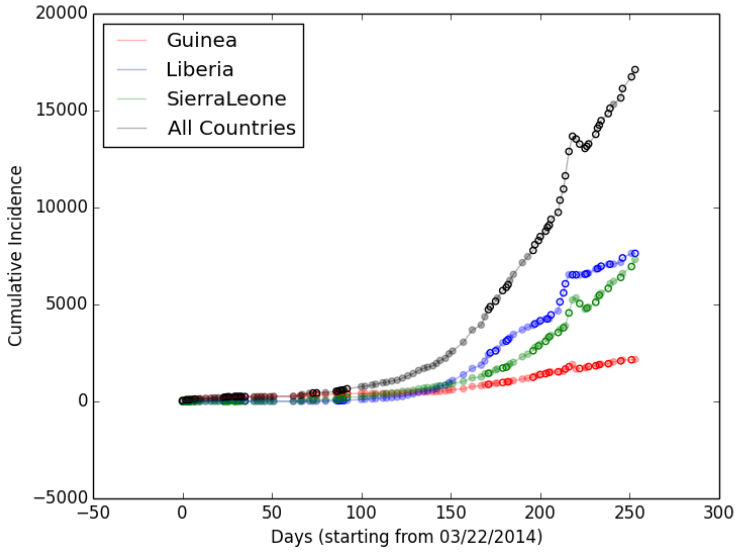


Figure 6: Data points obtained from local linear extrapolation points are circles, real data points are disks. For the "All Countries" curve, circles correspond to data points obtained from at least one "extrapolated" data point.

5.6 Outbreak Size

Total outbreak size can be obtained by integrating (1) such that:

$$I_{total} = \frac{exp\left(\frac{\ln(R_0)^2}{4\ln(1+d)}\sqrt{\frac{\pi}{\ln(1+d)}}\right)}{2}\sqrt{\ln(1+d)}[erf(x-\mu)-erf(-\mu)] \tag{5}$$

$$\text{Where } \mu = \frac{\ln R_0}{2\ln(1+d)}$$

The correctness of this formula is verified empirically in Figure 7

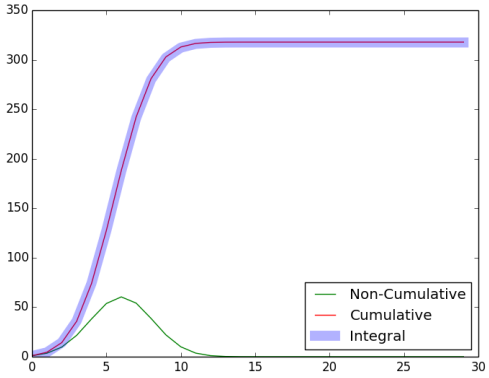
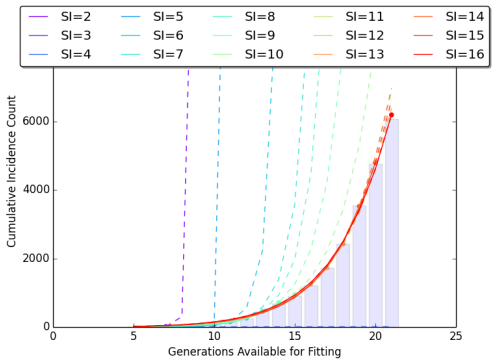


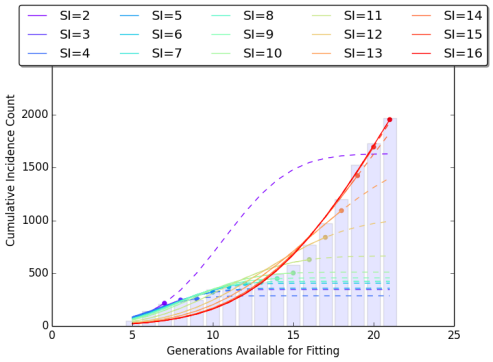
Figure 7: Discrete Cumulative Data at t vs Integration over t Simulated IDEA data for $R_0 = 3.91$ and $d = 0.12$

6 SUPPLEMENTAL DATA

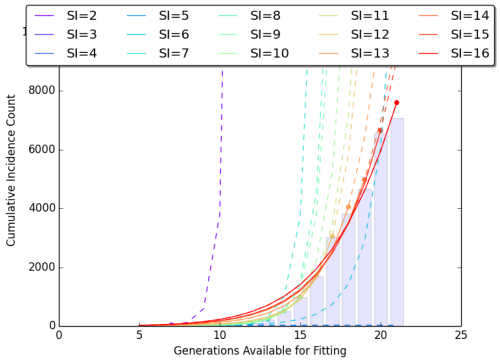
6.1 Country-Specific Progressive Projections



(a) Sierra Leone



(b) Guinea



(c) Liberia

Figure 8: Progressive Country-Specific Projections Projected epidemic growth curves based on generations available for fitting.

6.2 Country-Specific Sensitivity Analysis

Table 4: Guinea

	Ro	Ro (Fisman)	d	d (Fisman)
Base Case	1.693	2.007	0.013	0.027
12 day generation time	1.551	1.811	0.009	0.021
18 day generation time	1.853	2.170	0.017	0.032
Outbreak recognized generation 3	1.903	2.452	0.019	0.044
Outbreak recognized generation 7	1.545	1.733	0.008	0.017
Outbreak 50% underreported	3.506	5.350	0.066	0.143
Outbreak 99% underreported	4.049	3.651	-1.000	-1.000
Deaths only	1.631	1.965	0.012	0.029

Table 5: Liberia

	Ro	Ro (Fisman)	d	d (Fisman)
Base Case	1.714	1.232	0.009	-0.014
12 day generation time	1.626	1.190	0.008	-0.010
18 day generation time	1.978	1.301	0.015	-0.017
Outbreak recognized generation 3	1.949	1.423	0.015	-0.010
Outbreak recognized generation 7	1.550	1.107	0.005	-0.016
Outbreak 50% underreported	3.537	1.876	0.048	-0.039
Outbreak 99% underreported	4.002	3.538	-1.000	-1.000
Deaths only	1.654	1.147	0.009	-0.016

Table 6: SierraLeone

	Ro	Ro (Fisman)	d	d (Fisman)
Base Case	1.571	1.545	0.005	0.004
12 day generation time	1.505	1.446	0.005	0.003
18 day generation time	1.774	1.641	0.009	0.004
Outbreak recognized generation 3	1.763	1.821	0.009	0.012
Outbreak recognized generation 7	1.434	1.369	0.001	-0.001
Outbreak 50% underreported	2.959	3.023	0.030	0.034
Outbreak 99% underreported	4.001	3.595	-1.000	-1.000
Deaths only	1.501	1.387	0.007	0.001

REFERENCES

[1] 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. PubMed PMID:24391797.

[2] Fisman, David, Edwin Khoo, and Ashleigh Tuite. Early epidemic dynamics of the West African 2014 Ebola outbreak: estimates derived with a simple two-parameter model. Plos Currents Outbreaks (2014).

[3] Mercer, G. N., Glass, K., & Becker, N. G. (2011). Effective reproduction numbers are commonly overestimated early in a disease outbreak. Statistics in Medicine, 30(9), 984–94. doi:10.1002/sim.4174

[4] Ridenhour, B., Kowalik, J. M., & Shay, D. K. (2014). Unraveling Ro: considerations for public health applications. American Journal of Public Health, 104(2), e32–41. doi:10.2105/AJPH.2013.301704

[5] Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, et al. (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345: 1369–1372

[6] Arino, J., & Driessche, P. Van Den. (n.d.). The Basic Reproduction Number in a Multi-city Compartmental Epidemic Model, 135–142.

[7] Luksza, M., Bedford, T., & Lassig, M. (2014). Epidemiological and evolutionary analysis of the 2014 Ebola virus outbreak, (June), 1–19. Populations and Evolution. Retrieved from http://arxiv.org/abs/1411.1722

[8] Hufnagel, L., Brockmann, D., & Geisel, T. (2004). Forecast and control of epidemics in a globalized world, (Track II).