Reviewer’s report

Title: Modeling the variations in pediatric respiratory syncytial virus seasonal epidemics

Version: 1 Date: 29 April 2010

Reviewer: Erik Volz

Reviewer’s report:

The authors seek to explain the origin of biennial cycles in the severity of RSV epidemics among young children. They develop a model using ordinary differential equations and estimate several parameters of interest using time series of reported infections in a single medical center.

The questions addressed in this manuscript are interesting, but I am concerned with several choices regarding the statistical methodology. And several aspects of the ODE model are poorly justified.

1. Finding that early epidemic growth rates is correlated with final size is not at all surprising. But I find the analysis of this relationship inadequate. Is there a good reason to expect the relationship between growth rates and final size (or the other variables considered) to be linear? In SIR models of this sort, the final size (FS) is related to the Reproduction number by the self-consistent equation

\[ FS = 1 - \exp(-R_0 \times FS) \]

The exponential growth rate is approximately \( r = \beta - \gamma \), so \( r = \gamma \times (R_0 - 1) \) and \( R_0 = (r + \gamma) / \gamma \). You can then work out the relationship of FS with r, and it is *definitely* not a linear relationship.

If the authors choose to rely on a statistical analysis based on only 7 data points, I suggest choosing a parametric form that is inspired by the real mathematical relationship between FS and the growth rate. One might also vary the 4 week window to evaluate the robustness of the correlation between r and FS.

2. Another serious problem with this manuscript stems from the failure to incorporate births and age structure. Unfortunately, I think including these things may be essential in this case, since all of the reported cases occur in children under two years of age. Since an epidemic will last around 4-6 months, that means about a quarter of the
reportable susceptible population is turning over within each epidemic cycle, and
this can have a major impact on the results.

3. Methods/Model fitting, para 3: It is inaccurate to say that the authors used the
Bayesian Information Criterion, which assumes that proper likelihoods can be
calculated.
Instead, the authors are using a pseudo-likelihood based on a sum of squares.
The resulting information criterion is sometimes called PLIC (pseudo likelihood
information criterion).
Unfortunately, this raises the problem that the results reported in Table 3 may be
sensitive to the form of pseudo-likelihood used.
For example, would the same model be optimal if instead of the RSE
pseudo-likelihood, the authors used some other measure of deviance, like RMSE
?
Judging by Table 3, it seems possible that the model that allowed all parameters
to vary may be optimal, since it had close scores.
Pseudo-likelihoods can be used to compare models within an Approximate
Bayesian framework (see for example Toni & Stumpf, J of the Roy Soc Interface,
2009), which is, I think, a more principled approach than one using information
criterion.
I also think such an approach, in combination with sequential Monte Carlo or
MCMC, would be superior to the grid search employed here.
If the authors continue using PLIC, I think they must compare other forms of
pseudo likelihood for their conclusions to be convincing.

4. Conclusions, last paragraph, The authors enumerate a number of possible
extensions to the SEIDR model, including age stratification, births, and modeling
across many seasons with the incorporation of prior infections.
I already remarked that including births is essential.
And modeling multiple seasons with prior infections would make the main claim
of this paper much stronger-- that biennial cycles can be explained by partial
herd immunity from previous epidemic waves.
I don't see why these modifications should be left for future research-- they would
greatly strengthen the conclusions of this paper.

5. Figure 3, bottom left panel. The dashed(+) line should have allowed S0/N to
vary from year to year, but it appears flat.
I suspect the legend is incorrect, and the dotted triangle line actually represents
the model with variable S0/N (probably mixed up with the transmission
parameter).

6. I find the paper lacking details in how the grid search was conducted.
I think grid search should be fine if the pseudo-likelihood surface isn't too bumpy.
This is probably the case when fitting one parameter, though I have doubts about
the model with three free parameters.

How fine was the grid? How many parameter combinations were attempted? Is there any way to tell if the grid had sufficient resolution? Maybe examine how frequently the max-pseudo-likelihood is mis-specified as the resolution of the grid is varied.

- Minor Essential Revisions

1. Abstract, results, "exponential growth could predict epidemic characteristics"
This statement is much too strong; the authors have not shown that epidemic characteristics can be predicted with this method.

To do this properly would require the authors to divide the data into training and validation categories. But this is impractical considering that these regressions are based on only 7 data points.

The authors should simply state that epidemic growth rates are correlated with exponential growth rates.

Such language should also be corrected elsewhere in the document ("predict" has special meaning in statistics).

2. Methods / Regression analysis, para 1, How was the threshold of 4 weeks chosen?
There is a tradeoff-- as the interval on which rates are estimated increases, the estimate gains precision, but bias also increases since the epidemic trajectory deviates from exponential growth in the long run.

There should be some discussion of how this tradeoff was dealt with.

Ideally there would be a rigorous evaluation based on keeping bias within small bounds while maximizing precision.

3. para 2, Which correlation coefficient was used? Pearson?

4. Methods / SEIDR model, para 1, How does \( \alpha \) differ from \( t_0 \) used elsewhere in the text? It is claimed that \( \alpha \) was not "influential on the shape of the curve", but I do not see results to support this.

5. para 2, and the system of equations: Do these variables (S, E, I etc) describe the "fraction" of the population that is susceptibles, infecteds etc?
Or do they describe the absolute number of prevalent susceptibles, infecteds etc?

If the former, the system is frequency-dependant, and contact rates are independent of the number of infections.

If the latter, then this system of equations is density-dependant, and contact rates are not constant; that would be a somewhat unusual choice for an epidemic like this which should be justified.

6. para 2, Why do detected cases D transmit at the same rate as non-detected
cases I? I would suppose that once detected, an infected person would transmit
much less due to isolation and possibly treatment.

7. Methods / Model fitting, para 1, The definition of RSE seems a little ad hoc-- is
there a precedent?
A more natural measure of deviation might be the RMSE of standardized scores
\[ Z = \frac{x_i - \hat{x}_i}{x_i}, \]
and then
\[ \text{RMSE}(Z) = \sqrt{\frac{\text{sum} Z_i^2}{365}} \]
If variance of the x's is approximately proportional to size, the Z's might be close
to normally distributed.

- Discretionary Revisions

1. Background, para 1, should have citations for
several statements here (eg "treatment... improves overall health in the
population."). And why does cost-effectiveness of treatment increase if treatment
is only given during times of high RSV activity? To reduce wastage?

2. para 2, perhaps cite anecdotes as unpublished correspondence; this would
allow readers to follow up with the source.

3. para 3, citations for biennial cycle?

4. para 4, bottom, "variation in exponential growth could potentially be related to
variation in transmission rates, detection fractions, or proportions susceptible."
These are the variables the authors investigated, but there could be many other
important factors. Variations in network structure? It could also be affected by
the time of year when the epidemic begins.

5. para 6, I would enjoy a more detailed comparison to Weber's model. In exactly
which ways does your model differ? How would these differences affect your
results?

6. Methods / Data, para 3, Are results sensitive to the threshold of 5 cases for the
beginning of the RSV epidemic?
Would results be very different if it was 10 or 20?

7. Methods/ Regression, para 1, If \( \lambda \) was estimated on the basis of
prevalent infections (w/ could be estimated using known recovery rates) rather
than cumulative infections, then this rate could be related to R0. It would be nice
for the reader to report those estimates in addition to the rates.

8. Results/ SEIDR Model, last paragraph, I would put this in the Conclusions.

**Level of interest:** An article whose findings are important to those with closely
related research interests

**Quality of written English:** Acceptable
**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

i declare that i have no competing interests