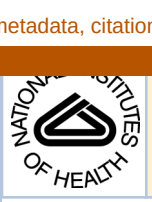

Published in final edited form as:

Epidemiology. 2012 September ; 23(5): 738–741. doi:10.1097/EDE.0b013e318258369b.

Visualizing Results from Infection Transmission Models: A Case Against "Confidence Intervals"

Eric T. Lofgren[1]

[1]Department of Epidemiology, UNC Gillings School of Global Public Health, UNC Chapel Hill, Chapel Hill, NC

Abstract

Stochastic transmission models are highly important in infectious disease epidemiology. The quantity of data produced by these models is challenging to display and communicate. A common approach is to display the model results in the familiar form of a mean or median and 95% interval, plotted over time. This approach has drawbacks, however, including the potential for ambiguity and misinterpretation of model results. Instead, we propose two alternative approaches for visualizing results from stochastic models. These proposed approaches convey the information provided by the median and 95% interval, as well as information about unexpected outcomes that may be of particular interest for stochastic epidemic models.

Increasing computational power has made it possible to use stochastic models to study the transmission of infectious diseases. These models, which include stochastic compartmental models, network models, and agent-based models, have appealing properties when compared with their deterministic analogs. A stochastic approach is advisable for small populations, as it builds results from the aggregation of individual-level data rather than population averages(which treat human beings as continuous rather than discrete units). Additionally, the use of stochastic models allows the examination of model outcomes that are unlikely but not impossible, outcomes that are ignored by their deterministic cousins.

Most stochastic models lack an approachable analytic solution, and rely instead on simulation of the model thousands of times. This leaves the researcher with the challenge of presenting tremendous amounts of data. One approach adapts the familiar technique of presenting the mean or median and 95% interval of the model results. While this presentation may be comfortingly familiar to most epidemiologists, it is problematic.

The confidence intervals produced by most statistical models have an agreed-upon meaning (albeit one occasionally subject to misinterpretation)and well-established methods for their construction. In contrast, transmission models rarely have analytic methods for producing confidence intervals. Most intervals produced by transmission models are not true confidence intervals but rather prediction intervals. Instead of being a summarization of the value of a true estimate based on repeated sampling, the intervals are a summarization of a large number of samples and the variability of those results. There is no clear, intuitive ordering of model results into percentiles, and different approaches may produce different interpretations of the same model. Furthermore, the results of many transmission models are multimodal, a property that is impossible to communicate with methods designed for

Mailing Address: Eric Lofgren, 4501 Connecticut Ave. NW Apt. 411, Washington, DC 20008, Phone: (303) 912-2595

Conflicts of Interest and Sources of Funding:
Eric T. Lofgren has no conflicts of interest to disclose, and is funded by NIH Training Grant 2T32AI070114

unimodal distributions. Finally, outlying results are often of interest in a transmission model and are hidden when using an interval-based approach.

Fortunately, common data-analysis software packages have graphical capabilities that ameliorate these problems, allowing the visualization of both "common" outcomes and outlying data by showing all results, rather than reductionist summaries. The limitations of summarizing results with an interval approach are explored here using an example network model. Two alternative methods are presented.

Methods

As an example, an outbreak was modeled across a social contact network, Goodreau's Faux Mesa High School. This network was simulated using an exponential random graph model fit to a real high school in the rural western United States using data from the National Longitudinal Study of Adolescent Health.[1–2] This analysis uses the largest connected component of that network, which contains 120 students linked by 188 mutual connections (Figure 1). The data for this network are available as part of the "statnet" R package.[3]

Disease transmission was simulated 1,000 times with a percolation model, using a single starting infected individual. The probability of transmission was set at 0.5053 to yield a basic reproductive number (R_0) of ~1.80. This value was chosen to distinguish the stochastic and deterministic forms of the same process. In a deterministic setting, this disease is guaranteed to cause an epidemic. However, in a stochastic setting the disease may become extinct before an epidemic begins.

The results of this model are presented using four approaches. In approach one, the median, 2.5^{th} and 97.5^{th} percentile of the total epidemic size were determined. The multiple realizations that resulted in those values were plotted over time. In approach two, the median, 2.5^{th} and 97.5^{th} percentile of currently infected persons were plotted against time. In contrast to approach one, these intervals do not represent a single realization of the model, but rather a pointwise summarization of the entire simulation at each time point. In approach three, all realizations of the model are simultaneously plotted as semi-transparent lines (to prevent overlapping trajectories from obscuring each other). Finally, in approach four, all realizations are plotted as points. The resulting scatterplot was smoothed using a kernel density estimator to highlight areas of dense results.

Simulations were performed with EpiFire 2.33(T Hladish, unpublished paper, 2011), and results were analyzed in R 2.13.[4] The results from our implementation, as well as the source code to produce the figures that follow, are available as electronic supplements (eAppendix 1 and eAppendix 2 respectively [https://github.com/elofgren/Visualizing-Results-from-Infection-Transmission-Models]).

Results

The epidemic simulation produced bimodal results, with 21% of realizations becoming extinct after the first infection, while the largest epidemic infected 87 people. The median epidemic size was 16 people, with any common measure of central tendency providing a poor summary of the results (Figure 2).

For approach one, the median epidemic size was 16 (with 2.5^{th} and 97.5^{th} percentiles of 0 and 66, respectively). Multiple realizations produced these values. Rather than make an arbitrary selection, four such realizations of the model for the median and both intervals are shown in Figure 3A. The graphical results show considerable heterogenity for the median

and upper interval. As their ordering is based on the total epidemic size, several median results appear above the upper interval results toward the beginning of the simulation.

For approach two, the median, 2.5th and 97.5th percentiles of infected persons are shown at each time point in Figure 3B. The median and upper-interval values do not represent a single particular realization. In many cases, the median and upper intervals take a non-integer value, which cannot represent an actual model outcome. In contrast, the lower interval represents a preponderance of model realizations. In approach two, intervals are extremely sensitive to whether epidemics that have died out are regarded as being zero infected people or as missing at time points past their extinction (not shown).

Approach three depicts all 1,000 model realizations(Figure 3C). The multi modal nature of the results is shown in the areas of dense dark lines near n=0 infections, as well as the modest epidemics ranging from 5 to 10 infections occurring from t = 2 to 10. Approach four similarly shows all 1,000 model realizations as a smoothed density (Figure 3D). Compared with approach three, the extreme results — while still apparent — are visually deemphasized.

Discussion

The importance of stochastic epidemic models will increase as computational resources become more available and simulations are used to address increasingly complex questions of infectious-disease transmission. Simultaneously, the ways that epidemiologists visualize these results will become more important, as these models elude compact numerical summarization.

Reductionist methods involving a median or mean and an upper and lower bound can be deceptive. While simple to digest, these plots hide potentially important information. In addition to discarding data, the composition of these plots can be somewhat arbitrary. The “median and 95% confidence interval” may refer to the final epidemic size, to each time point in the simulation, or to the midpoint of the simulation(among others). Additionally, each realization may be treated as a distinct trajectory or as a set of points to be averaged. These choices can affect the conclusions drawn from a plot.

Visualizations that capture the full details of the simulation accept increased complexity in exchange for retaining more information. Such visualizations do not require ambiguous summarization and have correspondingly less opportunity for misinterpretation. Approaches three and four are capable of displaying complex (e.g. multi-modal)results, while still identifying the most common results through visual density. These figures can be overlaid with numerical results or particular realizations of interest (eFigure [https://github.com/elofgren/Visualizing-Results-from-Infection-Transmission-Models]), allowing for faster interpretation without loss of information.

Whole-data approaches also preserve outlying results, which are often of great interest in transmission models. While these graphs require more initial explanation, the overall communication of results is better served with a plot appropriate to the nature of the data. Such graphs may be extended to other forms of analysis that generate large numbers of similar but distinct analyses, such as sensitivity analysis. Other approaches for representing the results of simulation models exist, including three-dimensional plots and animation. This paper illustrates two common but inappropriate techniques, and two informative alternatives that are tractable to code.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

I gratefully acknowledge the help and advice of Stephen R. Cole in the preparation of this manuscript.

References

1. Resnick MD, Bearman PS, Blum RW, et al. Protecting adolescents fromharm. Findings from the National Longitudinal Study on Adolescent Health. Journal of the American Medical Association. 1997; 278:823–32. [PubMed: 9293990]
2. Hunter DR, Goodreau SM, Handcock MS. Goodness of Fit of Social Network Models. Journal of the American Statistical Association. 2008; 103:248–258.
3. Handcock, MS.; Hunter, DR.; Butts, CT., et al. [Accessed October 15, 2011] statnet: Software tools for the Statistical Modeling of Network Data. 2003. http://statnetproject.org
4. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2008. URL http://www.R-project.org

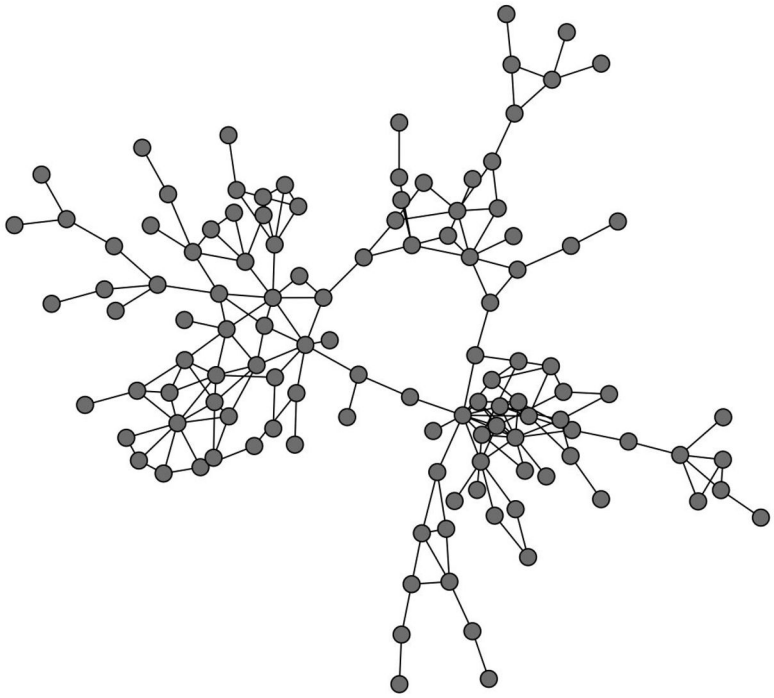

Figure 1.
Largest connected component of Goodreau's Faux Mesa High School social network. Each node (grey circle) represents a single individual, and edges(black lines)represent mutual connections.[1–2]

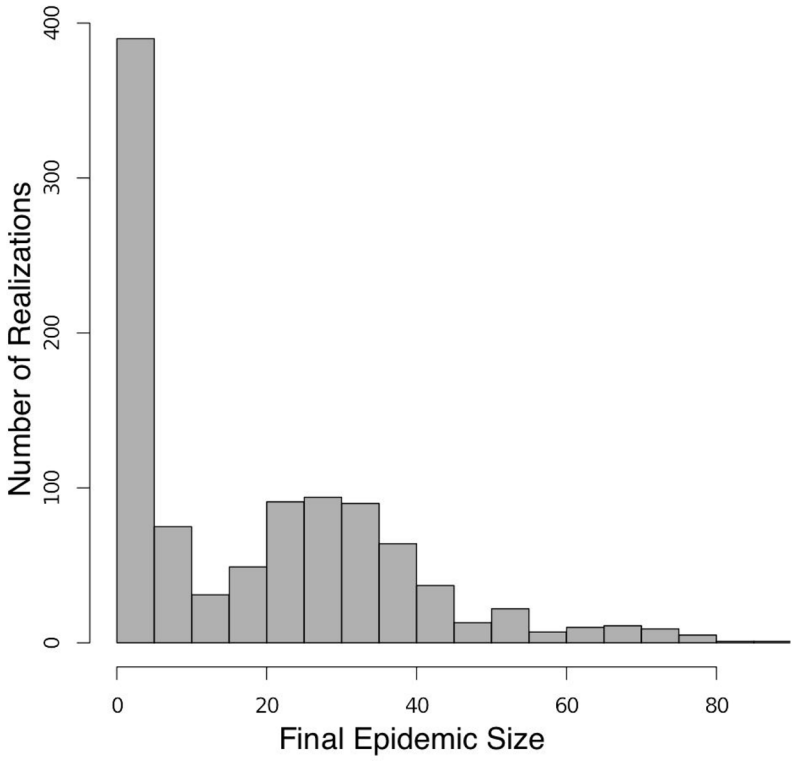

Figure 2.
Final epidemic size of 1,000 simulations of disease spread over a social network of 120 individuals with an R0 of 1.80. The most common epidemic size was 1 case (21%) with a median epidemic size of 16 cases.

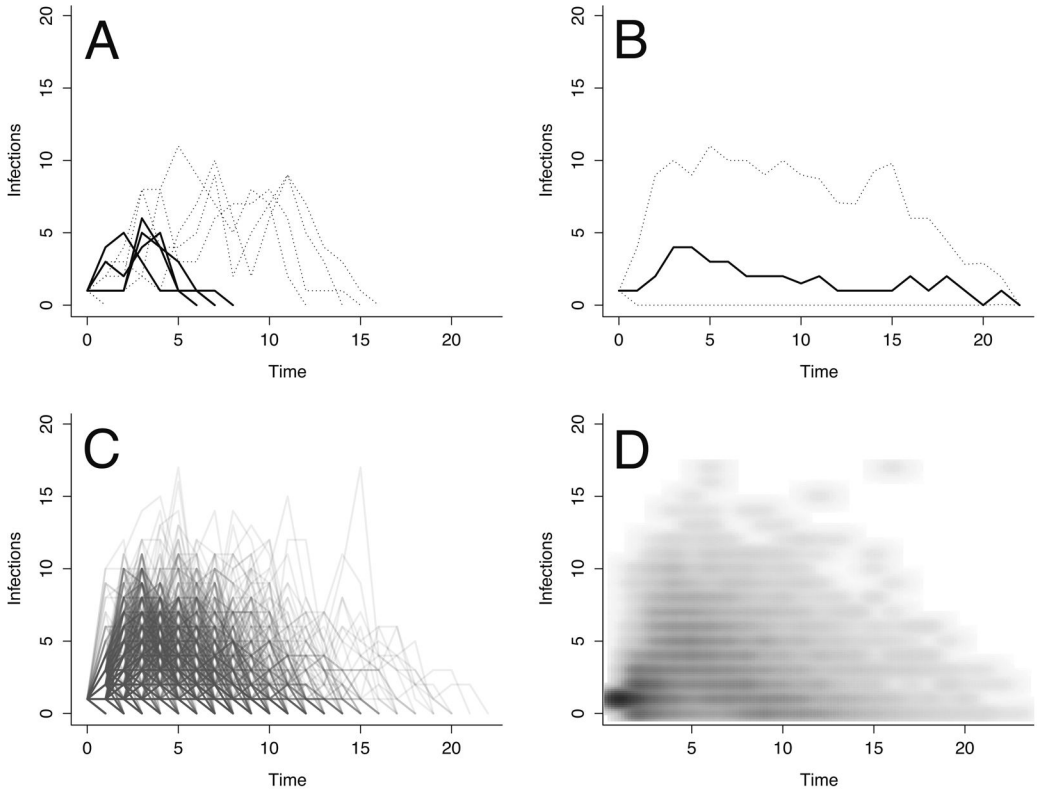

Figure 3.

Visual results of the simulated epidemic using approach one (A), approach two (B), approach three (C) and approach four (D). A, depicts the median and 95% confidence intervals based on the final size of the epidemic, with four choices for each as multiple realizations have the same final epidemic size. B, depicts the median and 95% confidence intervals based on the number of infected individuals at each time point. Epidemics that have experienced stochastic extinction are treated as missing, the default assumption in many statistics and graphing packages. C, shows the trajectories of all model realizations, with dark areas indicating more than one overlapping trajectory being simultaneously plotted. D, plots all model realizations as individual points and the resulting scatterplot smoothed using kernel density estimation. Darker areas represent correspondingly denser concentrations of simulation results.