Explicit Spatial-Temporal Simulation of a Rare Disease

Bian[1], D. Wilson[2], T Whalen[3], M Cohen[4], Y. Huang[1], G. Lee[1], E. Lim[1], L. Mao[1], Y. Yan

[1] State University of New York, Buffalo
[2] National Institutes of Health
[3] Georgia State University tom@whalen3.org (corresponding author)
[4] Frontline Healthcare Workers Safety Foundation

Abstract

This paper reports on the use of possibility theory and agent based explicit spatio-temporal simulation to compare the effects on each of three real communities given the assumption that a rare disease is carried out of a hypothetical high containment biological research laboratory sited in that community. The initial event has nonzero possibility but its probability is not well measurably different from zero. The conditional distributions obtained by making this event an input to the simulation resemble "counterfactual conditionals" that can provide useful information about the relative technical and social desirability of alternative sites even though a conventional risk assessment is not possible.

Keywords: Rare events, counterfactuals, geographic information systems, spatio-tempral models

1. Introduction

1.1. Rare Events, Rare Diseases

Release of an exotic disease such as Ebola from a high-containment biological research facility is an extremely rare event. In hundreds of thousands of person-hours of work in laboratories in the United States, there has not been even one clinical case of laboratory-acquired infection, let alone any cases of infection in the community. Such an event can be treated as an "adventitious" event [3] [9] [10].

Comparing the magnitudes of the minuscule threats to different local communities based on where a new facility is located can only be done in a relative manner by assuming, in effect, that the probability of a release is 100% when in fact we know that it is not measurably different from zero. Based on this counter-factual assumption, it becomes possible to compare the conditional probability of various degrees of harm taking the initial release as a given. This enables a possibilistic risk analysis to be done, even though a realistic probabilistic risk analysis would be impracticable because the point estimates would be substantially smaller than the width of the confidence intervals, and in many case the lower bound of the confidence interval is zero. Logically, the result may be considered a "counterfactual conditional." [7]

The agent based explicitly spatial and temporal modeling (A-BEST) approach used in this simulation is built on a conceptual framework developed by Bian [1] and an analytical framework by Bian and Liebner [2]. The approach uses the continuing advances in geographical information systems to quickly craft a high resolution, efficient model of disease transmission within a rural area, a small city, or a distinct community within a large city. This is useful for assessing risks associated with moderately communicable diseases such as those typically studied in high containment biological research facilities.

The simulation is constructed based on a hypothetical situation of a laboratory worker who sustains a laboratory-acquired infection such as Ebola despite all the equipment, procedures, and rules to prevent this. The worker is then assumed to leave the laboratory and enter the community without following any of the required precautions mandated after any possible exposure. Finally, he visits a prostitute during his convalescence despite standard warnings to abstain from intercourse during recovery from the disease, and infects her despite the fact that no known sexual transmission of actual Ebola has ever been observed. The simulation employs an individual-based and spatially explicit modeling approach for predicting possible health outcomes in three communities: urban, suburban==, and rural.

Many of the parameters of the disease model correspond to disease characteristics that are not precisely known. In the simulation setup, these uncertainties are resolved on a "worst case" basis. In some cases, such as the potential for sexual transmission, parameter values are used which go beyond what is justified by the data; "worse than worst case." Any reasonable variation in the estimates would increase the already high proportion of Monte Carlo runs in which there was no contagion at all.

1.2. Ebola

Ebola hemorrhagic fever (EHF) is a viral disease that occurs in humans and nonhuman primates. It is fatal in 50% to 90% of clinically ill cases. Researchers believe that the virus is zoonotic (animal-borne) and that it naturally occurs in an animal native to the African continent that does not get the disease.

Ebola virus can be transmitted through direct contact with blood, body fluids, or tissues of an infected person or animal, or contact with contaminated objects, such as needles. Spread of the disease through airborne particles has not been documented among humans in a real-world setting, such as a hospital or household. While virus or viral RNA has been found in seminal fluid in patients recovering from Ebola infection [4], there is no direct evidence of sexual transmission of Ebola virus [8]. Nevertheless, the US Centers for Disease Control and Prevention recommend a period of abstinence from unprotected sex during convalescence. [8]

Outside of two laboratory workers who became infected but did not transmit the disease to others [4] [6]. .

2. Population Simulation

2.1. Three Populations, Two Scales

The A-BEST model and subsequent simulations allow study of the interactions within a community that may lead to infectious disease transmission subsequent to a laboratory-acquired infection. The simulation focuses on the modeling of discrete individuals, individualized interactions, and how these interactions change with location and time [1] [2]. Based on this conceptual framework, the analytical framework further defines a multi-population and two-scale network. The two scales include a local-level and a population-level network.

The interactions between individuals at home, in workplaces, or at service places form a local network, whereas the mobility of individuals between these places links the local networks into the population level network. In this network, each individual is represented as a node and the interaction between any two individuals is a link. Collectively, these individuals and their individualized interactions portray a heterogeneous spread of disease through the three-population and two-scale network.

Three populations are modeled in this simulation, namely, a nighttime popula-

tion at home, a daytime population in workplaces, and a pastime population at service-places. Individuals within the populations travel between home, workplaces, and service-places. Disease can be transmitted from one location to another and spread through individual interactions consistent with daily life activities. [2]

These three populations represent the same set of individuals at different locations and time periods of a day. Each individual is assigned to a household, a workplace, and a number of service-places to create the three populations, and to establish the links between the populations. The spatial locations of homes, workplaces, and service-places and the travel routes for each individual between these three sets of locations are explicitly represented.

2.2 Simulating the Nighttime Population

Two sets of data are used to simulate the nighttime population at homes for the three communities, the 2000 census data and a household dataset purchased from ReferenceUSA, Inc. Since the US census statistics are created based on the residential locations of the population, they are used as the basis for the simulation of the nighttime population.

Two sets of information are extracted from the aggregated information, one for individuals and another for households in a block group. The information for individuals includes statistics of age range (children, adults, seniors), gender, and the relationship to the householder (householder self, spouse, own children, other children, other). The information for households includes statistics of household size and household type. The latter includes family households (married-couple, single-father, and single-mother families) and non-family households (living alone and not living alone).

2.3. Simulating the Daytime Population.

Three sets of information are used to simulate the daytime population at workplaces for the three communities, the 2000 census data, a "business" dataset purchased from Environment Systems Research Institute Inc. (ESRI), and the attributes associated with previously simulated individuals and households. While the ESRI data provide information about workplaces, the census data and the previously created attributes of the individuals and households help assign the individuals to these workplaces.

Collectively, these datasets help create the daytime population and establish the links between the nighttime population and the daytime population. The workplaces are defined as those businesses that are within a one-hour driving distance from the three communities. The driving distance is computed using Geographic Information Systems (GIS) software and the speed limit of the road network This definition results in a total of 310,400 workplaces. Each contains the following attributes: name, location, number of employees, estimated sales, and the type of business that is coded by the North America Industry Classification System.

Four sets of statistics information are extracted from the data:

(1) persons working outside the home,

(2) transportation means (walk, bus, car, and subway) used by these workers to travel to work,

(3) time needed for the workers' travel,

(4) type of employer (NAICS code).

The service-oriented workplaces are open in both daytime and pastime, thus two shifts of workers are assigned to these workplaces. For a given type of

worker and a given type of workplace, the assignment of a specific worker to a specific workplace is based on the Monte Carlo method. The school-aged individuals in each household are assigned to a school as their workplace.

This completes the creation of the daytime population at workplaces and links the nighttime population at homes to the daytime population in workplaces.

2.4. Simulating the Pastime Population.

The simulation of the pastime population at service-places is based on three sets of information. The first set is a 1991 travel diary survey data obtained from the Massachusetts Central Transportation Planning Staff. The second is a subset of workplaces that is identified as service-places. The third is the attributes for individuals and households created in the previous simulations for the nighttime and daytime populations.

The travel diary survey provides information on the type and frequency of services needed, the service-place data describe the characteristics of the service-places, and the attributes data help assign individuals to service-places. Subsequently, the three sets of data collectively help create the pastime population and link it with the nighttime and daytime populations.

A total of 89,159 workplaces are identified as service-places from the 310,400 workplaces based on their NAICS attributes. The individuals in households are assigned to service-places based on the attributes of individuals (gender and age range) and households (the number of workers in a household, the number of vehicles in a household, and the household size). Using the frequency statistics as constraints, the total number of trips for a household is first determined. Workers in a household, then other members of the household, are randomly assigned to one of the three types of service trips (workplace-to-service, home-to-service, and service-to-service), then to one of the six types of services. For a given type of service trip, a given type of service, and a given type of individual in a given type of household, the model assigns an individual to a nearby service place. Trips to service-places occur during both the daytime and pastime periods of the day.

Each service-place has two sets of individuals, those who need services and those who provide services.

3. Disease Transmission Model.

Using the A-BEST model, the transmission of a disease may follow any of an extremely large number of possible paths in the three-population and two-scale network.

After an initial infection ("index case") is introduced into a community, the model simulates the first generation of infection. That is, the model identifies the first set of individuals who are in direct contact with the first infected individual through contacts at home, in the workplace, and at service-places. Each of these first-set individuals is assigned an infection status, either healthy or infected, according to the primary infection rate of a specific disease.

(In the Ebola model, one simulated first-set individual is deterministically modeled as infected; as with the unrealistically high values of the parameters, this was done in order to have enough cases in each simulated community to allow comparison of counterfactual conditional infection rates.)

Next, the model simulates the second generation of infection by identifying the second set of individuals who are in direct contact with those infected in the first generation of infection. An infection status is assigned to these second-set indi-

viduals according to a secondary infection rate.

Thirdly, the model identifies the set of individuals who are in direct contact with those infected in the second generation of infection. These third-set individuals are assigned an infection status according to what is known as a tertiary infection rate.

The declining human-to-human infection rate between primary, secondary, and tertiary transmission is an important characteristic of zoonotic diseases. For example, most human cases of avian influenza were acquired from birds while a smaller number were acquired from humans who had been infected by contact with birds, but so far there are no cases of a human acquiring the disease from a human who had themselves acquired it from a third human.

The probability of an individual becoming infected after being in contact with an infected individual is based on the primary, secondary, and tertiary infection rate of a disease. For a given infection rate, a given disease, and a given set of individuals who are in direct contact with an infected individual, the Monte Carlo method assigns the infected status to certain individuals. Each infected individual experiences three stages of disease development: an incubation period, an infectious period, and (for survivors) an immune period. Infected individuals may recover or die based on the fatality rate of the simulated disease.

Table 1: Counterfactual Assumptions

1. High probability of sexual transmission even though none has been documented.
2. Transmission by casual contact even though only transmission by blood or other infected tissues or by contaminated medical equipment has been documented
3. Violation of multiple laboratory and CDC rules and procedures by index case.
4. Failure of hospital to place hemorrhagic fever patient in isolation.

Table 1 summarizes the counterfactual assumptions about the potential for contagion that were used to create a variant of Ebola which is possible in the sense of Lewis' "possible worlds" [Lewis, 2001] but whose probability is not well measurably different from zero.

4. Results

The results of the simulation include the day of the infection, the identification of the infected individuals, the time period and location of the infection, and the consequence of the infection (recovery or death). Results of each simulation overstate the actual risks to the communities, to a large degree, as a result of the highly unlikely initial scenario and the higher transmissibility of the simulated disease compared to Ebola.

Two hundred simulations in each community were conducted; 600 in all. Each of the 600 simulations included many thousands of interactions between simulated members of the populations.

Ebola Cases for Three Communities After Simulated Accidental Release.

	Infections		Infections per 10,000	
Community (Population)	(St. Er. Mean)	Mean	(st. Err. Mean)	Mean per 10,000
Suburban (30,022)	(0.06)	2.95	(0.02)	0.98
Rural (8,941)	(0.07)	3.07	(0.08)	3.43
Urban (284,805)	(0.09)	3.76	(0.003)	0.13

The data presented above clearly show that the presence of disease was not directly proportional to the number of people living in each community. The urban community has almost 15 times as many people, but only 1.25 times as many cases of disease on average than the average of the other two communities, and therefore far lower per-capita risk than the suburban or rural communities.

Close examination of the simulation outputs shows that the largest driver of the differences among localities is the size of the local hospital since this determines the number of health care workers who encounter the index case during the acute phase of the illness. In a real outbreak, this difference would be reduced or eliminated since any patient with a hemorrhagic fever would promptly be placed in isolation.

5. Summary and Conclusion

The continuing advances in simulation and geographical information systems make it possible for a multidisciplinary team to quickly craft a high resolution, efficient explicitly spatial and temporal model of a rural area, a small city, or a distinct community within a large city. This is useful for assessing risks associated with moderately communicable diseases such as those typically studied in high containment biological research facilities.

Release of an exotic disease such as Ebola from a high-containment biological research facility is an extremely rare event. Comparing the magnitudes of the minuscule threats to different local communities based on where a new facility is located can only be done in a relative manner by making counterfactual assumptions whose probability is not well measurably different from zero and evaluating the probability distribution of results conditional on those assumptions. This enables a possibilistic risk analysis to be done, even though a realistic probabilistic risk analysis would be impracticable

Many of the parameters of the disease model correspond to disease characteristics that are not precisely known. In the simulation setup, these uncertainties are resolved on a "worst case" basis. In some cases, such as the potential for sexual transmission, parameter values are used which go beyond what is justified by the data; "worse than worst case." Any reasonable variation in the estimates would increase the already high proportion of Monte Carlo runs in which there was no contagion at all. This might or might not be more realistic, but it would make it harder to discern differences in risk among the three communities.

References

[1] L Bian, "A conceptual framework for an individual-based spatially explicit epidemiological model," *Environment and Planning B, 31:*3, pp. 381-395, 2004.

[2] L. Bian and D. I-Iebner, "A Network Model for Dispersion of Communicable Diseases." *Trans in GIS. 11*:2, pp. 155-173, 2007.

[3] M. Cohen and T. Whalen, "Implications of Low Level Human Exposure to Respirable B. anthracis." *Applied Biosafety 12:*2, pp. 109-115, 2007.

[4] R.T. Emond, B. Evans and E. T. Bowen., "A Case of Ebola Virus Infection," *Brit Med J. 2(6086)*, pp. 541-544, 1977.

[5] D. Heymann, *Control of Communicable Diseases Manual,* American Public Health Association, 2004.

[6] ITAR-TASS News Agency, "Research scientist dies of Ebola fever in lab accident," (in Russian), Fri 21 May 2004.

[7] D, Lewis, *Counterfactuals (2Rev Ed),* Wiley-Blackwell, 2001.

[8] A. K. Rowe et al, "Clinical, virologic, and immunologic follow-up of convalescent Ebola hemorrhagic fever patients and their household contacts, Kikwit, Democratic Republic of the Congo," *J Inject Dis. Supplement 179:528535*, 1999,

[9] T. Whalen and C Bronn, "Possibilistic Risk Assessment," *Procs N Amer Fuzzy Information Proc Society Conf,* New York, 2008.

[10] T. Whalen, T. Taylor and M. Cohen, "Modeling Nearly Impossible Hazards," *Proc, 11th Joint Conf on Information Systems*, 2008.