Statistical Methods for the Classification of Partially Observed Outbreaks

Cosmin Safta, Jaideep Ray
Sandia National Laboratories, Livermore, CA
Karen Cheng, David Crary
Health Effects and Medical Response Group, Applied Research Associates, Inc., Arlington, VA

OBJECTIVE

The paper describes a novel statistical method to identify the etiological agent of an ongoing outbreak caused by an attack with an aerosolized pathogen. It is based on the fact that the time-evolution of an epidemic in a population, and the symptoms observed, are characteristic of the pathogen in question.

BACKGROUND

The identification of the etiological agent of an outbreak is typically performed via diagnostic tests conducted on patients. The time delay waiting for diagnosis can result in disastrous outcomes. In some cases, access to the patients or the necessary apparatus may also be restricted. Nevertheless, it is often possible to follow the evolution of the outbreak as a time-series of reported morbidity. Identification of the causative agent and estimation of the attack parameters (i.e., the number of people infected, the time of infection and a representative dose) are critical to predicting future levels of morbidity and consequently the degradation of mission capabilities. In principle, automated outbreak classification would reduce the time delay prior to crisis response.

The identification of the causes of the outbreak poses some significant challenges. The observations are typically noisy, as they include background morbidity. Further, there exists a stochastic time delay between the appearance of symptoms and their reporting; thus, the reported / observed values from the recent past are merely a lower bound on the actual levels of morbidity. Since one has a complete time-history of reports, it may be possible to dynamically estimate the reporting delay and thus "correct" the observations, which can then be used to estimate the attack parameters [1]. This calculation can be performed for all competing etiological agents; hypothesis testing using likelihood ratios or posterior predictive tests can be used to select the agent (model) most likely to reproduce the observations.

In this paper we demonstrate the methodology on a test problem. Inhalational anthrax and plague are proposed as competing etiological agents for an observed time-series of non-specific ILI symptoms with significant respiratory distress.

METHOD

We simulate an inhalational anthrax attack and "corrupt" the symptomatic patient stream with a randomly chosen reporting delay. The time-history of observations are used to estimate a "correction factor" for the number of symptomatic people. This correction is based on the fact that the observed number of symptomatics on a given day asymptotes to the correct number as the observation period increases. These "corrected" observations are used to estimate the attack parameters, using both Bayesian methods [1] and Maximum Likelihood Estimation (MLE). This is performed under the assumption that the morbidity is caused by anthrax and plague, the two competing hypotheses. Identification of the pathogen is performed via both a likelihood-ratio test and Bayesian model selection. These tests measure the ability of the inferred attack parameters to reproduce the observations for the two competing agent models.

RESULTS

Preliminary results show that the MLE estimates of attack parameters are very sensitive to the starting guess, while the Bayesian method is robust. Further, model selection in case of a low dose anthrax attack (mean dose $<10^3$ spore per person) can be performed reliably, partially because of the significant difference between the incubation periods of the two diseases at low doses. Tests involving higher doses and small populations (<5000 infected) are ongoing.

CONCLUSIONS

Preliminary results indicate that it may be possible to identify/shortlist the causative agents of an outbreak from partial observations. Our method is very sensitive to the completeness of the set of competing hypotheses. The method is sufficiently accurate to supply parameters to epidemiological models in order to evaluate the demand on a medical system, but should not be used to determine the medical treatments since very different pathogens may have similar epidemiological signatures.

REFERENCES

1. J. Ray et al, Sandia Technical Report SAND2008-6044.

Corresponding author: C. Safta, csafta@sandia.gov

Presented at the 8th ISDS Conference, Miami, December 3-4, 2009.